

Multiple protein feature prediction with statistical relational learning

Luca Masera

October 1, 2016

Chapter 1

Introduction

1.1 Motivations

Proteins are the workhorse molecules of life, they take part in almost everything what happens in and outside the cells, from building structures such as hair or nails, to enzymes, which accelerate chemical reactions. It is sufficient to consider some number to understand their importance and variety: the total number of proteins in human cells is estimated to be between 250,000 to one million and the dry weight of our bodies is made by about the 75% of proteins [40]. Hence, it is easy to understand that an accurate analysis of their features and properties is crucial to understand the deepest mechanism of life.

Since their identification in the eighteenth century, proteins have been a focal point of micro-biologist researches, but it is just after the half of the twentieth century that the study of proteins makes its biggest progresses. Indeed, in that years Sanger accurately determined the correct amino acid sequence of insulin, and Max Perutz and Sir John Cowdery Kendrew respectively resolved the three dimensional structure of hemoglobin and myoglobin through X-ray crystallography. These discoveries opened the floodgates to a completely new set of protein analysis techniques. In the 80's indeed, bi-

ologist recognized the potential of the firsts portable computer and began to store their newly obtained protein sequences and structures data in digital format. Since their marriage between computer science and biology has been long and fruitful.

Protein sequencing however proceeds much faster than the ability of wet biologists to annotate the genes product. Moreover computer science technological breakthroughs and next generation sequencing techniques exacerbate this difference. Indeed, at the beginning of 2015 there are almost 200 millions¹ genes sequenced and only the 0.25%² of them have been manually curated and annotated. These statistics clarify that new and reliable computational approaches are needed to fill this gap and in this field machine learning techniques find breeding ground for their application to biological problems. In fact in the last decades they have been widely applied to the protein feature prediction task, but their results are yet not precise enough to guarantee reliable prediction to biologists. This research area is therefore nowadays very hot and alive.

1.2 Contributions

The aim of this work is to face the protein feature prediction, both from a statistical and a relational point of view. This bioinformatics task is a multi-labeled problem where it is needed to assign multiple annotations to a single protein. However these labels are not completely independent one from the other and therefore knowledge on the relations between them could positively impact on the quality of the predictions.

In standard machine learning algorithms indeed, knowledge lies only into the dataset and the usual way to enrich it with field-specific background information is via sophisticated kernel functions. Unfortunately this opera-

¹<http://www.ncbi.nlm.nih.gov/genbank/statistics>

²<http://web.expasy.org/docs/relnotes/relstat.html>

tion is not always effective, it is indeed hard to formulate relations between learned classes. Semantic-Based Regularization (SBR) [8], a state-of-the-art relation statistical learning framework developed in the University of Siena, provides a possible solution to this lack allowing the supervisor to instill background information into the well established kernel machine learning process in form of first order logic constraints.

SBR gives us the tools to integrate both aspects and achieve better protein feature prediction. In particular, we focus our attention primarily on the prediction of the Gene Ontology annotation.

From the relational point of view the aim was to develop first order logic constraints that express knowledge on the relationships between the predicted labels. The first set of logical rules has been obtained from the consistency property of Gene Ontology, i.e. if an annotation is assigned to a gene product (in our case proteins), then all the annotation ancestors in the hierarchy must be associated with that gene product. The second set of information exploited in this work comes from a protein-protein interaction dataset, whence we derived additional rules to submit to our framework.

In addition to the effort done in generating valid and informative rules that could effectively improve the learning process, we implemented four kernel functions that exploited different aspect of the protein, going from the simple amino acidic sequence to the expression levels of the genes responsible for their creation.

1.3 Structure of the document

Background

Our study is about a very specific field, i.e. prediction of protein function, so in the first part background knowledge are presented and defined to give the reader the basic tools to understand the subject treated in this document. At the beginning we will illustrate the importance of proteins, which

play a crucial role in our work, describing their structure and highlighting the relevance of these bio-molecules for life. Subsequently the main bioinformatics resources for retrieving (e.g. protein databases and gene ontology) and analyzing (e.g. sequence alignment algorithms) protein information are provided. Moreover, it is taken an overview of the literature about the several ways how this yet hard task has been faced until now. The considerations made in these sections will also be useful to understand the choices made in the kernel functions selection and the biological conclusions in the results.

After the biological excursus, the machine learning tools used in this work are presented and illustrated. After the explanation of the features and theoretical firmness of kernel machines, the document comes to the description of SBR, the relational statistical learning framework exploited in the protein feature prediction task.

Methods

The purpose of this chapter is to explain the method developed in the experimental phase of the work. After the formalization of the our task, it is provided the description of the kernel functions used, which will have an important part in the results discussion, before proposing the developed rules to inject prior knowledge into the learning process. The methods chapter ends with the description of the pipeline implemented to automatize the generation and execution of experiments and the metrics used to evaluate them.

Results

Since this work is an experimental one, this chapter covers the most important part of the document, explaining the results obtained with our relational statistical learning setting on more then 1500 *S. cerevisiae*'s proteins. The discussion is divided based on the type of feature analyzed and will be focused on the effects of kernel functions and set of rules used.

Conclusion

The conclusions of this document are drawn in this final chapter, in which the goals that we achieved are presented and explained, correlated to the baseline purposes of our work and the novelties found during the developing of it. It will be specially focused on the confirmed and disproved hypothesis that we set at the very first beginning. Moreover will be also found useful ideas for future works on this topic, that could further been explored.

Chapter 2

Biological background

2.1 Proteins

This work is mostly centered around the prediction of protein features. It is therefore important to give the reader an overview on the main concept related to the argument and the taste of the importance of the role that proteins play for life.

Proteins are biological macromolecules composed of one or more amino acids chains and, together with lipids, sugars and nucleic acids, are considered to be the fundamental building blocks of life. Examples of important biological roles played by proteins may include catalytic activity (e.g. amylase contribute to digestion), signaling and regulation (e.g. somatropin also known as growth hormone stimulates growth and cell reproduction in humans and other animals), molecule transport (e.g. Na^+/K^+ -ATPase allows the exchange of sodium and potassium in neural cells), structural support (e.g. tubulin, the monomer of the microtubules in the cytoskeleton) and movement (e.g. myosin allows muscle contraction).

The biological information needed for generating new copies of proteins is contained in the genes, that are DNA portions. The process by which a gene is copied into a RNA filament is called transcription. The transcribed

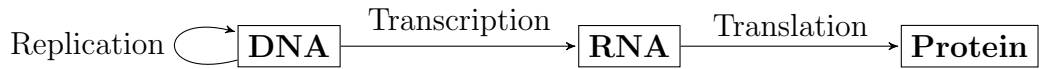


Figure 2.1: **Central Dogma of Molecular Biology.** *The Central Dogma of Molecular Biology describes the three fundamental processes for life. Is interesting to notice that all these are carried out by proteins.*

RNA, called messenger RNA (mRNA), is then synthesized into a amino-acid chain by the ribosome, which is itself a protein (with some RNA chains). This step is called translation because each triplet of nucleotides in the mRNA is translated into one of the 22 amino acids and linked together to form a new protein. Transcription and translation are crucial for life, so much that together with the DNA replication, which is also mediated by proteins, form the central dogma of molecular biology (Figure 2.1).

2.1.1 Structure

Proteins are long polymers and amino acids are their monomers. Amino acids are organic molecules composed of a common backbone, consisting of amine ($-\text{NH}_2$) and carboxylic acid ($-\text{COOH}$) functional groups, along with a specific side-chain, that characterize each amino acid. Indeed, the physico-chemical properties of the side chain determine the structure and the shape of the whole protein.

After the translation, proteins fold themselves in complex three dimensional structures, which will widely determine their functionalities. During the folding process, not only the amino acids composition is important, but also the environment plays a critical role. Indeed, according to it, some amino acids may prefer to face the outside of the protein or the core. If the folding does not succeed the protein is useless or, in some cases, even dangerous. It is indeed known that many degenerative diseases like the Alzheimer, the Cystic fibrosis or the BSE (commonly known as "mad cow" disease) are caused by the misfolding of proteins. Therefore, there exist specialized proteins (called molecular chaperones) and biological process (translocation)

that try to guarantee to proteins the right folding environment.

The structure of a protein is usually analyzed at four different levels (Figure 2.2), which are:

- Primary structure: is the lowest of the four levels and is determined by the bare sequence of amino acids. Nowadays it is reasonable easy to obtain this information by translating the nucleotide triplets in coding sequences of genes obtained by genome sequencing, but is less informative than the higher levels.
- Secondary structure: represents how the peptidic backbone interacts with itself through hydrogen bonds. At this level specific folding shapes can be identified, i.e. the alpha helix and the beta sheet. This first level of folding occurs just after the translation (few milliseconds) and is mainly caused by hydrophobic behaviors of the amino acids.
- Tertiary structure: describes how the secondary structures interact with themselves forming the three dimensional shape. The structure is stabilized through hydrogen bonds and disulfide bridges (covalent bonds created post-translation). Information relatively about this levels are still hard and expensive to obtain. Bioinformatic methods are not powerful enough to face this problem, so techniques like NMR (Nuclear Magnetic Resonance) or X-ray crystallography have to be applied.
- Quaternary structure: some proteins are constituted by more than one peptidic chain or are part of protein complexes. These interactions are described by the quaternary structure. Studying this structure level is usually even harder than for the tertiary one. Due to the steric size of protein complexes, they can not be easily crystallized and analyzed with X-ray crystallography. In this cases in then used the cryo-electron

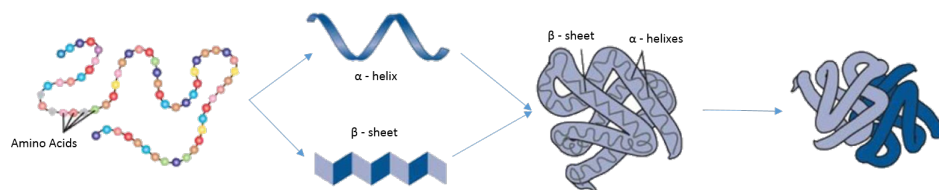


Figure 2.2: **Protein structure.** In figure are shown the schematic representation of the four levels of the protein structure.

microscopy technique that allows to literally freeze the structure of a protein and investigate it through an electronic microscopy.

Even after having finished the folding process, the structure of a protein is not rigid and immutable. Their three dimensional structure can indeed change in time depending on the surrounding environment (temperature, pH, voltage, ion concentration, phosphorylation or ligand binding), which can actually trigger their functionalities. It is, for example, the case of myosin which interacts with ATP molecule. ATP molecules are very energetic and their hydrolysis induces a conformational change in the protein that, as a final result, allows the muscle contraction.

Evolution has deeply shaped living beings, giving them very different forms and features, from their appearance to their molecular level. Proteins are no exception. Despite this there are portion of them, which are strongly conserved between different proteins of the same organism but also between species. This pieces of proteins are called domains and are usually characterized by a compact structure and a specific folding. Protein domains are strongly related to their function and are combined in proteins like building blocks. Their conservation can be used as key for identifying known domains in proteins according to their primary sequence.

2.2 Bioinformatic resources for proteins

In the first years of the 80's two technological breakthroughs were gaining a foothold in the scientific world: the Sanger sequencing methods and the first computers. The former made it possible to acquire rapidly and accurately (for the time) the sequence of long stretches of DNA, the latter to store and analyze this information. Since then the marriage between biology and computer science has been long and successful. Many specific algorithms and resources have been developed by computer scientists and then applied by computational biologists to explore the deepest aspects of life. In this section are proposed the most relevant results of this fruitful marriage with a special focus on the protein analysis.

2.2.1 Algorithms

Because there are only 22 amino acids¹ that constitute proteins, proteins primary structure is usually represented as a string using an alphabet of 22 symbols $\Sigma = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, U, V, W, Y\}$. This makes proteins very suitable to be treated with string algorithm to detect similarities that can be related to functional or structural relationships.

When protein sequences are compared, scoring function that reflect biological or statistical relevance of a couple of amino acids are employed. So, instead of using a binary function (1 if considering the same amino acid, 0 otherwise), the similarity score is obtained by substitution matrices like PAM (Point Accepted Mutation) [7] or BLOSUM (Blocks Substitution Matrix) [18]. The first one expresses the evolutionary acceptability of a specific point mutation, the second instead encodes the empirical probability of an amino acid. Scoring functions are also able to deal with gaps in sequences, assigning them specific penalties based on position and length.

¹Selenocysteine and Pyrrolysine are often not included in the list because very rare and until a few years ago considered of secondary importance.

Sequence alignment algorithms are a vast class of algorithms whose goal is to identify the "best" possible between two sequences. The first distinction that can be done is in the definition of "best" alignment. There are indeed two different approaches to the problem, i.e. the local and the global. Global methods give more importance the total number of matching into the sequences. This can obviously leads to a fragmented solution that do not take into account the modular nature of proteins. Differently, local alignment algorithms attempt to maximize the alignment piece-wise. This approach is very useful when comparing dissimilar sequences in order to identify conserved regions such as motifs or domains. A further differentiation involves the number of sequence treated per-time by the algorithm.

Pairwise alignment methods are efficient algorithms that consider two sequences at a time and could be local or global. They are usually used to deal with similarity searches in databases or to extrapolate information about an unknown sequence from the characteristics of a known one. The three primary pairwise alignment methods are:

- Dot-matrix method: is an exhaustive algorithm that tries all possible alignments between two sequences. It builds a matrix D , where the entries represent the quality of the alignment near to that position.

```
>gi|129295|sp|P01013|OVAX_CHICK GENE X PROTEIN
(OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPPHVTQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLMLVLLPDEVSDLERIEKTINFEKLTWETNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESQFRADHP
FLFLIKHNPTNTIVYFGRYWSP
```

Figure 2.3: **Example of amino acidic sequence in FASTA format.** *This is an example of a primary sequence retrievable in one of the many primary databases. The ">" symbol introduces the description line, a series of meta information used to identify the sequence. The following sequence is then formatted in such a way that each line of text is shorter than 80 characters in length.*

It can be formalized as follows: given a sliding window w , which represents the number of consecutive amino acids to average, and two sequences s and t , the entry in position i, j of the dot-matrix is equal to $D_{i,j} = \frac{1}{w} \sum_{k=0 \dots w} S(s_{i+k}, t_{i+k})$ where S is the scoring function used (e.g. binary, BLOSUM62, PAM120, ...). The out-coming matrix can be easily plotted mapping the values of each entry into a gray scale (an example in Figure 2.4). If the data are not too noisy and the size of the sliding window is correctly selected, repeated portions of the sequences and gaps can be clearly observed. However, this method can be mainly appreciated graphically because it is hard to extract match summary statistics and match positions of two sequences.

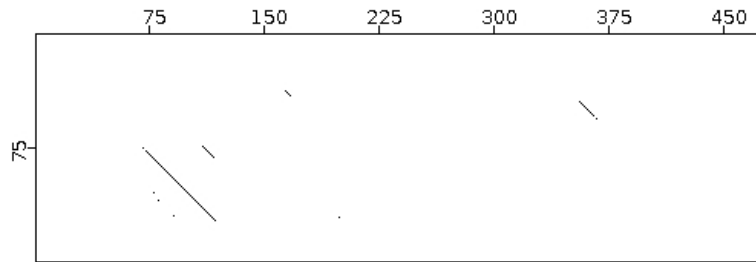


Figure 2.4: **Dot-plot of two *S. cerevisiae* proteins.** The plot shows the results of the dot matrix method on two *S. cerevisiae* proteins, i.e. Q00776 and Q00381. A highly conserved region can be noticed. This corresponds indeed to a shared protein domain that is the AP complex, *mu/sigma* subunit.

- Dynamic programming methods: are a class of algorithms that guarantees to find the optimum alignment, both in the local (Smith-Waterman algorithm) and in the global (Needleman-Wunsch algorithm) case. Given two sequences of size n, m the two algorithms generate a $n \times m$ matrix that is progressively (dynamic programming) filled. Each entry represents the score of the best possible alignment up to that position. The difference between the two methods lies in the choice of the value to assign to the i, j position, but, once the matrix is completed can reconstruct the final solution starting from the last entry in the matrix

and going backwards up the first one.

- Word methods: they are heuristic methods that, differently from dynamic programming ones, are not guaranteed to reach a global optimum, but are significantly faster. The key idea of these class of methods is to use k -tuple (sequences of length k) to identify regions that will probably be in the final alignment, and from these starting the refinement. These methods often rely on substitution matrices for scoring pairs of amino acids and apply to gaps negative scores. The correct parameters of the matrix and of the size of k may highly impact on the final results, but the choice is often more empirical than theoretical. The two most famous implementations of this class of algorithms are FASTA [31] and BLAST [1], which are nowadays still widely used.

*One or two homologous sequences whisper,
a full multiple alignment shouts out loud.*

(Hubbard et al.,1996)

Multiple alignment Multiple alignment (MSA) techniques attempt to align three or more sequences at the same time. These methods are often used to identify conserved region across a group of proteins highlighting phylogenetic relationships between them. They are, however, computationally very intensive² and heuristics are necessary for keeping them feasible.

- Dynamic programming methods: similarly to the pairwise case, dynamic programming can be applied to discover the optimal multiple alignment between k sequences. However, in the MSA case, instead of building a two dimensional matrix, it requires a k -dimensional matrix,

²Most formulation of the multiple alignment problem lead to NP-complete combinatorial optimization problems [46]

but the concept is analogous. The matrix is filled progressively and the score of each cell is computed according to the previously computed ones. Unfortunately the complexity is exponential in the number of sequences and therefore the problem becomes quickly intractable for increasing values of k .

- Progressive alignment construction methods: are heuristic methods whose approach consists in two main steps. In the first one the algorithm generate a tree, called *guide tree*, where two node are neighbor if they have the best pairwise alignment among all pairs. Then, starting from the couple of sequences with the highest similarity score, the final solution is obtained by incrementally adding new sequences according to the *guide tree*. The most popular method of this class is the ClustalW algorithm [45].

2.2.2 Databases

The huge amount of biological data produced with the new sequencing methodologies has largely overtaken the biologist capability to analyze them manually, therefore new computational techniques are needed not only for mining informations, but also for storing and organizing them. In the last decades a vast variety of biological databases (BDB) has therefore appeared allowing easy storage and consultation of data. According to their content, databanks can be classified as primary, the one storing primary sequences, and derived, which contain informations obtained by the analysis of primary sequences.

Important protein related BDB are:

- UniProt (The universal protein resource) [4]: is the biggest bioinformatics database created by the European Bioinformatic Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). It collects protein sequences of most of

living beings and viruses from the main publicly available databases and organize them in a comprehensive, non-redundant database (UniParc). Moreover part of this sequences are manually curated by experts. The annotation, such as protein function, subcellular localization, protein protein interaction domains and active sites are stored in the UniProtKB/Swiss-Prot database.

- Protein Data Bank (PDB) [3]: is the main repository for retrieving the three-dimensional structure of proteins and nucleic acids and is therefore a key resource of structural biology. Each entry in the database consists of all coordinates of all atoms in a specific protein. This data are submitted by biologists and biochemists from all over the world and typically obtained with X-ray crystallography or NMR spectroscopy. Stored information can be retrieved in a text format, but also as 3D browsable images. Since its foundation in 1973, the number of analyzed proteins is highly increased reaching at the beginning of 2015 almost 100000 proteins structures.
- CATH [15, 42]: is, similarly to SCOP, a hierarchical classification of protein domains. The name is itself the acronym of the four classification levels:
 - Class: classifies the domains according to their overall secondary structures composition (mainly alpha, mainly beta, alpha beta and few secondary structures).
 - Architecture: takes into account the general spatial arrangement of secondary structures of the domain.
 - Topology: is analogous to architecture, but including connectivity between secondary structures.
 - Homologous superfamily: this classification is manually curated and highlights an evolutionary relationship (at least two criteria

from sequence, structure or function must be observed).

- Structural Classification of Proteins (SCOP)[28]: is biological database containing a structural classification of protein domains. Its aim is to provide an evolutionary relations between proteins. At a first look the levels of SCOP (Class, Fold, Superfamily, Family etc) are similar to the CATH ones, but there are some classification differences.

2.3 Gene Ontology

The lack of a standard terminology in the biological field can lead to inefficient communications and ambiguous data sharing in the scientific community. Gene Ontology (GO) [2] is a bioinformatics project -part of ‘The OBO Foundry’- whose purpose is to fill this lack providing a controlled vocabulary of univocally defined terms representing gene product properties of all species.

2.3.1 The structure

Gene Ontology is a manually curated structured ontology with the aim to provide a unique and unequivocal description of all gene products. This information consist of an ID, a name, the GO domain and a natural language definition listing its main features. All the entries are organized in three direct acyclic graphs (DAGs), each one having as root a specific GO domain, which are:

- cellular component: the parts of a cell or its extra-cellular environment,
- molecular function: the elemental activities of a gene product at the molecular level, such as binding or catalysis,
- biological process: operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living

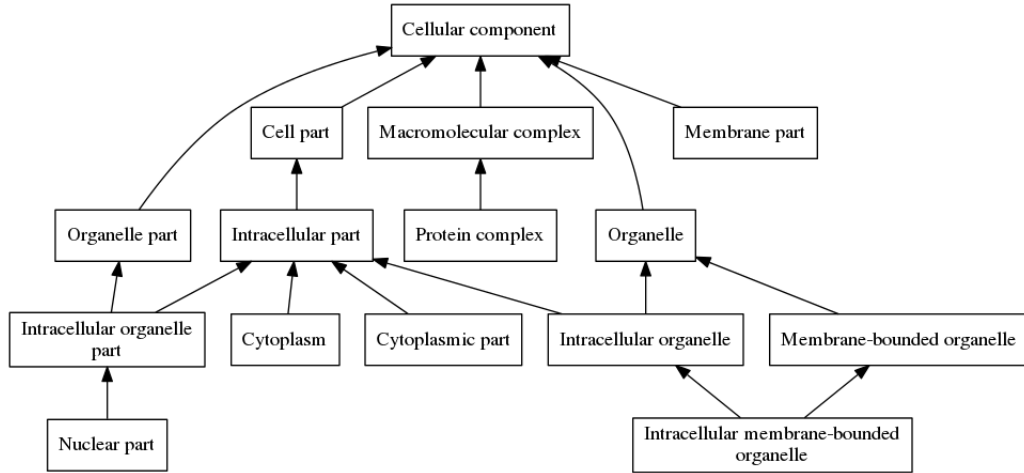
units: cells, tissues, organs and organisms.

In addition to the annotation, relation plays a very important role in the GO structure definition, giving fundamental information about relationships between the various terms.

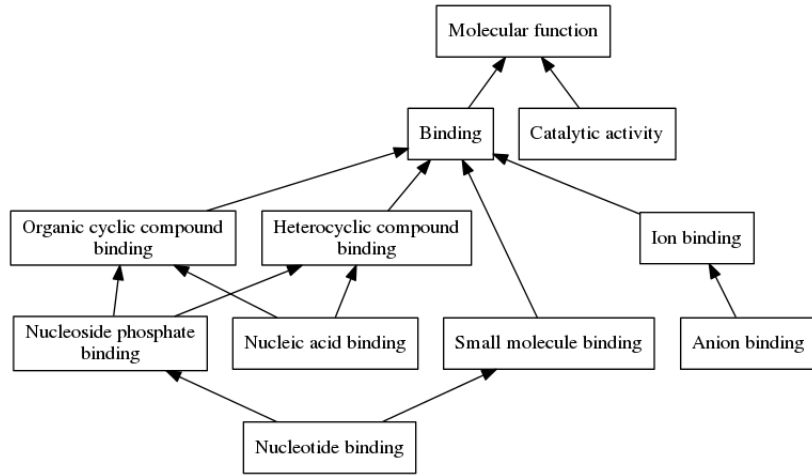
- *is a*: is the main relation and forms the basic structure of GO. A GO term A is said to be in a *is a* relation with a GO term B if A is a subtype of B. Therefore $A \Rightarrow B$, but not vice-versa. This relation is transitive, indeed taking the *is a* chain of mitochondrion, intra-cellular organelle and organelle as example, it is clear that a mitochondrion is an organelle and not every organelle is mitochondrion.
- *part of*: is similar to the *is a* relation, but it express the concept of being part of something (cellular component) or take part in something (biological process and molecular function). As the *is a* relation, *part of* is transitive, indeed if A *part of* B and B *part of* C, then A *part of* C. This relation can be trans-hierarchy, e.g. a Molecular Function can be *part of* a Biological Process.
- *regulates*: this relation represents the ability of a GO term to directly affects the manifestation of another. It can be further specified if the regulation is positive (at the growth of the first the second grows) or negatively (at the growth of the first the second decreases).
- *occurs in*: this relation expresses a locational relationship between a BP and a CC term. It is probably the most informative trans-hierarchy relation, but unfortunately it occurs rarely in the GO dags.

A fundamental property considered in this work, which computational models often ignore and also neglect, is what in [25] is called the True Positive Rule (TPR). This property derives directly by the *is a* relation and guarantees that if a protein belongs to specific GO term, then it belongs

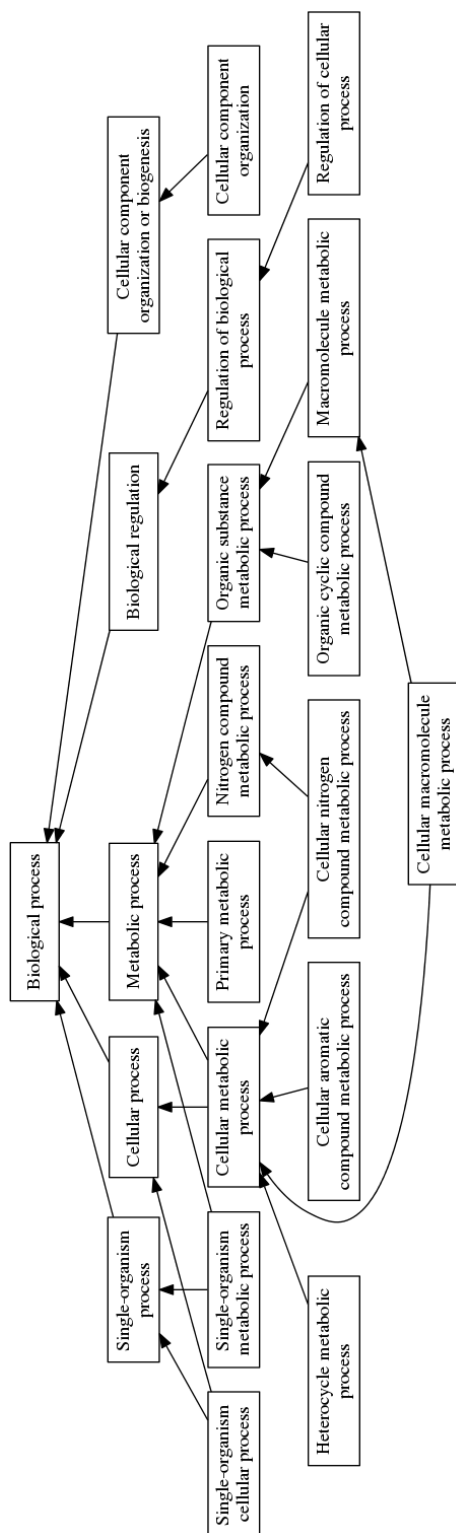
also to each node in each path from that term to the root if *is a* relations are followed. This property can be translated into consistency constraints as shown in Section 4.3.



(a) Cellular Component



(b) Molecular Function



(a) Biological Process

Figure 2.4: **Gene Ontology DAGs.** *In this figure are reported three sub-graphs of the first levels of Gene Ontology*

2.4 Bioinformatics approaches to the protein feature prediction problem

Literature concerning protein function prediction is really vast and the spectrum of methodologies even more. Attempts to identify the state of the art in this fields led to the CAFA³ (Critical Assessment of Function Annotation experiment) challenge project, a worldwide effort aimed at analyzing and evaluating protein function prediction methods. A summary of the first competition can be found in [35] and points out that the abundance of biological data positively impacted on the results, showing big improvement with respect to traditional approaches like BLAST [1], but there is no clear dominant methodology that overtakes the others.

The goal of this section is to highlight the major classes of approaches used to face this problem. The methodologies taken into account are most relevant appeared in the CAFA competition whose details are extracted from an exhaustive survey done by [30].

Sequence-based methods

Homology based

In the lately 80's the first sequence similarity systems like FASTA [31] and BLAST [1] make their appearance. Since then they have been widely applied to analyze nucleotidic and aminoacidic sequences from the major databases to find homologies in proteins.

Homology is a evolutionary concept that means some trait of two organisms derive from a common ancestor. Applying this concept to proteins we obtain that two proteins are homologous if they derive from a common ancestor protein. The idea behind comparing amino acidic sequences is that highly conserved sequences may maintains their functionalities. Nonetheless two proteins deriving from a common ancestor (very similar sequence) may

³<http://biofunctionprediction.org/>

evolve different functions. This kind of proteins are called paralogous. It is the case for example of the Gal1 and Gal3 proteins of the *S. cerevisiae*. They have an identity 385 amino acids over 529 and a similarity of 443 amino acids over 529 but they cover completely different roles in the cell: Gal1 is a galactokinase (part of a metabolic pathway) while Gal3 is a transcriptional inducer [33].

Motif based

Motif-based methods are another sequence-based protein function annotation method. These techniques do not compare complete protein sequences. The investigated sequence is compared with known motifs, which are shorter signatures for protein domains, and for each of them is provided a statistic significance level. The commonly used measure is the Expected value or E-value and the P-value. The former represent the number of matches with the similarity score obtained one can expect to find by chance in sequence of that size. The more it is close to zero, the more significant the match is. The latter instead measures the probability to find a matching with that similarity score by chance. The most popular database for retrieving specific information for motifs are probably Prosite [41] and InterPro [21]. Protein domains are well known to be strongly related to protein functions, especially at molecular level. Therefore a statistical significant match probably implies that the protein owns that particular function.

Moreover motifs can be used to predict the cellular localization of a protein. During the translocation phase, special peptidic sequences are used to identify proteins and target them to their folding location. This information can be used as hint for their final sub cellular localization.

Structure based

Protein structure is often more conserved then the sequence, but there are evidences that highly similar structures may show different functions. There are for example 27 homologous protein superfamilies folding in a *TIM*

barrel shape that cover over 60 *EC numbers*⁴ [15]. There are some explanations to this aspect, the first is purely technical. Three dimensional protein structures are mainly obtain through X-ray crystallography, which requires, like the name says, a crystallized protein. Depending on the substrate in which the protein usually is folded, the crystallization may considerably alter the original shape of the protein. The second one is that, especially in catalytic activity, not only the shape of the active site is important but also the physico-chemical composition of the residues in the active site is crucial.

Nonetheless proteins with highly similar structures are likely to have the same or similar function. On this assumption rely many popular methods like *DALI* [19], *SSAP* [20], *STRUCTURAL* and *CATHEDRAL* [23].

Protein Interaction Networks

Protein tends to perform their functions together with other proteins. It is often exactly this interaction that allows complex biological process. For example, myosin proteins by themselves would not be able to contract any muscle without the substrate made by the actin and this is just one of the many several examples of cooperation between proteins. This trend of proteins to co-work inspired some protein function prediction methods. Protein-Protein Interaction (PPI) networks are usually represented as graphs where proteins are the nodes and relationships the edges.

The underlying idea of this methods is that proteins that lie close in the PPI network are likely to share functions (especially at the biological process level). Indeed once the topology of the network is known, function annotations can be propagated to their neighbors or to the cluster they belong [38]. Unfortunately PPI networks are not easy to obtain and make these methods feasible just for well studies organism where the topology is at least partially known.

⁴Enzyme Commission number is a numerical classification of enzymes, based on the reaction they catalyze.

Gene expression data

Gene expression experiments are a quantitative measure of the transcription rate of the mRNA during the protein synthesis. The most common technique to obtain information about genes expression level in a specific biological situation is the microarray analysis. Microarrays are silicon plates on which tens of thousands of probes are attached, where the probes are nucleotidic sequences corresponding to a specific gene⁵ which will interact with a marked cDNA⁶. Comparing data obtained by the analysis of different samples in various biological conditions, co-expression data can be inferred.

Assuming that co-expressed genes are likely to be biologically related and have part to the same biological process, the natural approach is to cluster them according to their expression profiles. Indeed, genes related to the same biological process are expected to be transcribed and silenced together. Function of the unannotated proteins are assigned according to the dominant function in which they are. This process can be summarized in the *guilt-by-association* concept.

Multiple Data Type

Combining multiple data sources is a widely used technique in both, informatics and biology. The union of different information origins can improve the quality of the single dataset, indeed the errors of one can be compensated from another. Moreover data sources covering different aspects of the subject give a more comprehensive context to the analysis.

The conjunction of the different information can lead to a common format for all of them or to independent ones. An example of the former is the multiple kernel learning, where different kernel are used to build a final one in which sub-kernel are weighted according to their informativeness.

⁵Probes have an high affinity to the target gene (complementary sequence of the mRNA), but in some cases also non-perfect matching could occur resulting in false positives.

⁶The mRNA sequences extracted from the cell are treated with the reverse transcriptase, which is an enzyme that generate a complementary DNA (cDNA) from an RNA template.

Regarding the former, this work is a perfect example. Indeed the knowledge coming from the kernels is combined with the one of the constraints whose information origin is completely different.

Chapter 3

Methodological background

3.1 Kernel methods

Kernel methods are a class of machine learning algorithms for pattern analysis, whose best known exponents are the Support Vector Machines (SVM) [5]. Thanks to their theoretical firmness, their computational efficiency and flexibility, kernel methods are ubiquitously presents, not only in the machine learning literature, but also in fields like computational biology where they are often used as black boxes to solve the most diverse tasks.

The common ground among these methods is the use of *kernel functions*. These functions allow to operate in a high-dimensionality feature space, without the cost of computing the explicit inner product between feature vectors. This means that *kernel functions* merge the expressiveness of non linear machine learning algorithm with the efficiency of the linear ones. Intuitively *kernel functions* generalize the notion of dot product to arbitrary (even infinite-dimensional) input space and can be seen as a measure of similarity between objects. The flexibility and expressiveness of *kernel functions* arise from the fact that they can be reasonably easy applied not only to point in the euclidean space, but also to data structures, such as graphs, trees or sequences (extremely interesting in bioinformatics).

Definition 3.1 (Valid kernel function [37]) Given a function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$ and a set of examples $\{x_1, \dots, x_n\}$, the Gram matrix $K \in \mathbb{R}^{n \times n}$ is defined as:

$$K_{i,j} := k(x_i, x_j) \quad (3.1)$$

If the Gram matrix generated by k is positive semi-definite, i.e. if it satisfies the condition:

$$\sum_{i,j} c_i c_j K_{i,j}, \quad \forall \mathbf{c} \in \mathbb{R}^n \quad (3.2)$$

then k is a valid kernel function.

Valid *kernel functions* are very important in machine learning because they always correspond to a dot product in some Reproducing Kernel Hilbert Space (RKHS). Moreover the Representer Theorem [37] shows that problems in the form:

$$f^* = \arg \min_{f \in \mathcal{H}} c((x_1, y_1, f(x_1)), \dots, (x_n, y_n, f(x_n))) + g(\|f\|) \quad (3.3)$$

where \mathcal{H} is an appropriate RKHS, g a strictly monotonically increasing real-valued function and c an arbitrary cost function, admits representation in terms of kernel expansions of the form:

$$f^*(\mathbf{x}) = \sum_{i=1}^n \alpha_i k(\mathbf{x}, \mathbf{x}_i) \quad (3.4)$$

where $\alpha_i \in \mathbb{R}$ for all $i \in \{1, \dots, n\}$. The significance of this theorem is that it demonstrates that many learning methods allow solution that can be expressed as expansion in the training data.

Example (Support Vector Machines[5]) Given the labeled dataset $D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ with $y_i \in \{+1, -1\}$, the SVM classifier (without the

bias¹ [9]) can be interpreted as a regularization method in which:

$$c((\mathbf{x}_1, y_1, f(\mathbf{x}_1)), \dots, (\mathbf{x}_n, y_n, f(\mathbf{x}_n))) = \gamma \sum_{i=1}^n \max\{0, 1 - y_i \mathbf{w}^T \Phi(\mathbf{x}_i)\}, \quad (3.5)$$

and $g(\|f\|) = \frac{1}{2}\|w\|^2$. $\Phi(\cdot)$ is the function mapping the input example in some feature space. By letting $\gamma \rightarrow \infty$ we obtain the Hard-margin version.

Focusing on this last point we obtain:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{subject to:} \quad & y_i \mathbf{w}^T \Phi(\mathbf{x}_i) \geq 1 \quad \text{for all } i \in 1, \dots, n \end{aligned} \quad (3.6)$$

that is exactly the primal formulation of the hard-margin SVM. The constrain derives by infinitely penalizing misclassified samples. Consider now the dual formulation of the problem:

$$\begin{aligned} \max_{\lambda \in \mathbb{R}^n} \quad & \sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j y_i y_j \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j) \\ \text{subject to:} \quad & \lambda_i \geq 0 \quad \text{for all } i \in \{1, \dots, n\} \\ & \sum_{i=1}^n \lambda_i y_i = 0 \end{aligned} \quad (3.7)$$

where λ_i for all $i \in \{1, \dots, n\}$ are the lagrangian multipliers. The instances for which $\lambda_i \alpha_i > 0$ are the *support vectors*. In this formulation the decision function will be in the form: $f(\mathbf{x}) = \sum_{i=0}^n \lambda_i y_i \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x})$. Data examples appear just in the form of $\Phi(\mathbf{x})^T \Phi(\mathbf{x}')$ and can therefore be replaced by the *kernel function* $k(\mathbf{x}, \mathbf{x}')$. This technique is known as "*kernel trick*" and leads

¹The biased version of the Representer theorem do not change the significance of the obtainable results, but adds technicalities

to this kernel formulation:

$$\begin{aligned}
& \max_{\lambda \in \mathbb{R}^n} \quad \sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \\
& \text{subject to: } \lambda_i \geq 0 \quad \text{for all } i \in \{1, \dots, n\} \\
& \quad \sum_{i=1}^n \lambda_i y_i = 0
\end{aligned} \tag{3.8}$$

In this case the decision function could be expressed as:

$$f(\mathbf{x}) = \sum_{i=1}^n \alpha_i k(\mathbf{x}, \mathbf{x}_i) \tag{3.9}$$

where $\alpha_i = \lambda_i y_i$.

■

3.1.1 Kernels in bioinformatics

Computational biology and bioinformatics communities have extremely benefited from the introduction of kernel methods, developing on top of them a multitude of kernel functions from scratch or adapting already existing ones. In the following will be given some examples of kernel functions that has been successfully applied to bioinformatics tasks.

Spectrum and mismatch string kernels

Spectrum string kernel functions, also known as k -mers kernels, are a family of kernel thought to compare sequences on the base of the number of their common contiguous sub-strings of length k (k -mers). These methods have been proposed by Leslie et al. in [27, 26].

Given a sequence \mathcal{S} from an alphabet \mathcal{A} of size $|\mathcal{A}| = l$ and a number $k \geq 1$, the k -spectrum of \mathcal{S} is the set of all the subsequence of length k that it contains. Let $\Phi_k : \mathcal{X} \rightarrow \mathbb{R}^{l^k}$ be the function mapping \mathcal{S} in its feature space (all possible sequences of length k from the alphabet \mathcal{A}). Then:

$$\Phi_k(\mathcal{S}) = [\phi_\alpha(\mathcal{S})]_{\alpha \in \mathcal{A}} \tag{3.10}$$

where $\phi_\alpha(\mathcal{S})$ is the number of times α occurs in \mathcal{S} . The k-spectrum kernel for sequences \mathcal{S}_1 and \mathcal{S}_2 is then

$$K_k(\mathcal{S}_1, \mathcal{S}_2) = \langle \Phi_k(\mathcal{S}_1), \Phi_k(\mathcal{S}_2) \rangle \quad (3.11)$$

Depending on the size of the alphabet and the value of k the vectors out coming from $\Phi_k(\mathcal{S})$ can be very sparse and therefore the dot product for computing the kernel very expensive. In practice is however not necessary to compute neither the explicit vectors nor the explicit dot product. Indeed, in [27] is proposed an efficient computation of the kernel based on suffix trees whose overall complexity is $O(n \log(n))$ where n is the length of the longest sequence.

In [26] Leslie et al. propose an alternative version of this kernel, the mismatch string kernel. This kernel expand the concept of string kernel by allowing up to m mismatches when comparing two.

3.2 Semantic-Based Regularization

Semantic-Based Regularization (SBR) is framework developed by [8] at the University of Siena, whose development stems from the need to incorporate prior knowledge into the well established machinery of statistical learning algorithms. The previous standard approach to introduce field-specific knowledge into the learning process was mainly focused on building sophisticated kernel functions, that could express the more information as possible. Unfortunately many interesting relationships between learned categories could not be easily incorporated into a kernel function. SBR approaches this problem allowing the user to inject prior information, in the form of First order Logic (FOL), that describes the relationships occurring between and inside the classes.

Semantic-based regularization, as the name may suggest, is formulated

as multitask regularization problem on top of which is added a penalty cost coming from the infringement of the constraints (semantic part) as shown in Equation 3.15. The main feature of SBR are described in [8] and can be summarized in:

- *Learning from constraints in kernel machines*

The usual kernel methods machinery has been expanded in order to integrate the new concept of constraints, whose satisfaction degree is assessed through unsupervised data. This assessment allows SBR to refine his predictions trying to balance the label error of the training set and the constraints violation of the unsupervised examples.

- *Bridging logic and kernel machines*

T-norms allows the translation of first order logic clauses in real-valued functions that take part to the learning phase, ending up with a constrained multi-task optimization problem. The conversion of logic clauses into real values builds a natural bridge between logic and kernel machines.

- *Stage-based learning*

Unfortunately, unlike classical kernel methods, constraints expressed in logic clauses increase the computational complexity, ending up in some cases with non-convex optimization problem. This makes typical optimization approaches not suitable. Therefore [8], inspired by the work of Jean Piaget on stages of cognitive development, developed a stage-based gradient descent algorithm. This approach consists essentially in two phases. In the first one only supervised example are taken into account running the learning process until convergence without constraints, which are added only during the second phase. Being supervised example, at least in theory, coherent with the constraints, the first stage gives better starting point, rather than a random one, for

the seconds stage gradient descending.

The ability to merge FOL constraints and multi-task kernel machines, makes SBR very attractive for bioinformatics studies. It is the case of "Improved multi-level protein-protein interaction prediction with semantic-based regularization" of [36]. In this work constraints are used to improve the protein-protein interaction by imposing the need of two compatible domains (one per protein) and at least five interacting residues per protein. These boundaries refined the prediction increasing the consistency of the results. These are great hypothesis also for this work because they allow us to represent grounded biological information obtained from Gene Ontology as FOL rules achieving the consistency deriving from the True Path Rule. The actual constraints are explained in Methods.

3.2.1 Theoretical foundation of SBR

Let us consider a multitask learning problem, with T the total number of tasks and where each k -th task is represented by a function f_k defined in an appropriate Reproducing Kernel Hilbert Space (RKHS) \mathcal{H}_k . In [8] is proposed a more general version of the Representer Theorem that extend it to multitask optimization problems. Indeed Equation 3.3 can be generalize to:

$$[f_1^*, \dots, f_T^*] = \underset{f_1 \in \mathcal{H}_1, \dots, f_T \in \mathcal{H}_T}{\operatorname{argmin}} E[f_1, \dots, f_T] \quad (3.12)$$

and each function in the solution can be expressed in the form:

$$f_k^*(\mathbf{x}_k) = \sum_{x_k^i \in S_k} \alpha_{k,i}^* K_k(\mathbf{x}_k, \mathbf{x}_k^i) \quad (3.13)$$

where K_k is the kernel corresponding the the space \mathcal{H}_k and S_k the set of available samples for the task k .

Given that the tasks have to satisfy a set of H constraints the assumption made in SBR is that these task are correlated. The constrains are then

defined by the functionals:

$$\phi_h : \mathcal{H}_1 \times \dots \times \mathcal{H}_T \rightarrow [0, +\infty) \quad (3.14)$$

and $\phi_h([f_1, \dots, f_T]) = 0$ for $h \in \{1, \dots, H\}$ (i.e. the constraints are satisfied) must hold for any valid choice of $f_k \in \mathcal{H}_k$ for $k \in \{1, \dots, T\}$. In Subsection 3.2.3 we will show how the constraints are actually translated into real values.

The resulting optimization problem can be now expressed as a combination three component (Equation 3.15), the first penalizing the complexity of functions, the second penalizing the errors in the training set and the last one penalizing the infringement of the constraints.

$$\lambda_R \sum_{k=1}^T \|f_k\|^2 + \sum_{k=1}^T \sum_{(x_k^i, y_k^i) \in \mathcal{L}_k} L(f_k(x_k^i), y_k^i) + \lambda_C \sum_{h=1}^H \phi_h(\mathcal{S}, f) \quad (3.15)$$

where L is a loss function for the labeled examples and \mathcal{S} is the set of examples available. The meta-parameters λ_R and λ_C impact respectively on the contribute of the regularization part and of the constraints one. Representer theorem allows us to rewrite this optimization problem in term of kernel expansion. Let $f_k = \mathbf{G}_k \boldsymbol{\alpha}_k$, then we obtain:

$$\lambda_R \sum_{k=1}^T \boldsymbol{\alpha}_k^T \mathbf{G}_k \boldsymbol{\alpha}_k + \sum_{k=1}^T L(\mathbf{G}_k \boldsymbol{\alpha}_k, \mathbf{y}_k) + \lambda_C \sum_{h=1}^H \phi_h(\mathbf{G}_1 \boldsymbol{\alpha}_1, \dots, \mathbf{G}_T \boldsymbol{\alpha}_T) \quad (3.16)$$

where \mathbf{G}_k and $\boldsymbol{\alpha}_k$ are respectively the gram matrix and the weight vector for the task k .

The optimization is now in term of the weights vectors and can be done by descending the gradient. As described at the beginning of this section, SBR introduces a stage-based learning procedure by which the initial steps of the gradient descent are done with $\lambda_C = 0$ in order to find a local minimum from where introducing the constraints. This heuristic is necessary because

constraints in most interesting cases (such as this work) are non-linear and make therefore the problem non-convex.

3.2.2 First Order Logic: an overview

First Order Logic (FOL) is symbolical reasoning language in which statements are composed by logical concept (predicates) associated with logical subjects (variables) and quantifiers. It finds his main application field in the formal reasoning. FOL is a syntactical and semantical expansion of the proposition logic.

Propositional Logic (PL) is a relatively simple yet expressive formal language based on the concept of propositions, which given a truth assignment can be true or false. Propositions are logical formulas whose building blocks are proposition variables and logical connectives that are, in order of priority, \neg (not), \wedge (and), \vee (or), \Rightarrow (imply) and \Leftrightarrow (double implication or equivalence). The table 3.1 shows how the logical symbols modifies the values of the logical formulas A and B .

Propositional logic performs extremely well for expressing specific concepts, but lacks of the possibility to bound a variable to a concept. For example the statement ‘*Sara is happy, then she smiles*’ can be easily expressed in PL (*sara.is.happy* \Rightarrow *sara.smiles*), but statements like ‘*There is a person named Sara*’ can not be expressed in a compact and significative way in PL. First Order Logic extends propositional Logic with the aim to allow the association of a subject to a concept and their quantification.

Firs order logic inherits the logical symbols of PL to which it adds the

A	B	$\neg A$	$A \wedge B$	$A \vee B$	$A \Rightarrow B$	$A \Leftrightarrow B$
T	T	F	T	T	T	T
T	F	F	F	T	F	F
F	T	T	F	T	F	F
F	F	T	F	F	F	T

Table 3.1: **Logical connectives truth table.**

quantifiers \forall (for all) and \exists (exists) and an infinite set of logical variables x_1, x_2, \dots . In addition to the logical symbols, FOL uses logical predicate $P(\cdot)$ which returns if the predicate is satisfied given the input. We are now able to express the statement ‘*There is a person named Sara*’ in FOL as $\exists x. Person(x) \wedge Named_Sara(x)$ where x is a logic variable, and $Person$ and $Named_Sara$ are two logic predicate with arity one (number of parameters). If a grounding of the variable x is given it is now possible to state whether the statement is satisfied.

First Order Logic is a powerful tool that will allow us to write complex biological relations in compact and unequivocal way.

3.2.3 Translation of First Order Logic into real valued constraints

The translation of a First Order Logic formula into a real value is fundamental for converting the constraints into a penalizing term in the optimization problem. Without loss of generality the FOL formula can be rewritten moving all the quantifiers at its beginning (Prenex Normal Form).

$$\underbrace{\forall x \forall y}_{\text{Quantifiers}} \cdot \underbrace{(A(x) \wedge A(y)) \Rightarrow (B(x) \wedge B(y))}_{\text{Quantifiers-free formula}} \quad (3.17)$$

It can be noticed that given a grounding of the logic variables, the quantifiers-free part of the formula is a formula in propositional logic and can be therefore mapped in $[0, 1]$ with one of the methodologies coming from the context of *Fuzzy Logic*.

Fuzzy Logic and T-norms

Fuzzy Logic is an extension of the Boolean logic in which the concept of true and false are ‘blurred’. Indeed to statements are assigned truth values ranging in $[0, 1]$, where 0 is completely false and 1 completely true.

In order to operate in this logic the usual operands have to be reformulated to be able to handle real values. Notice that as in the boolean logic

only two operators are necessary to obtain all the others. In this case we will use \neg and \wedge . Let $N : [0, 1] \rightarrow [0, 1]$ and $T : [0, 1] \times [0, 1] \rightarrow [0, 1]$ respectively the mapping of \neg and \wedge . Then:

$$N(x_1) = 1 - x_1 \quad (3.18)$$

where x_1 is a fuzzy logic variable. It is clear that when x_1 approximates 0 or 1 the result is exactly what we would expect from the boolean logic.

The function T generalizes the conjunction and correspond to a family of functions called *T-norms* (triangular norms). *T-norms* must satisfy the following properties: commutativity ($T(x_1, x_2) = T(x_2, x_1)$), monotonicity ($T(x_1, x_2) \leq T(x_2, x_3)$ if $x_1 \leq x_2$ and $x_2 \leq x_3$), associativity: ($T(x_1, T(x_2, x_3)) = T(T(x_1, x_2), x_3)$) and the number 1 acts as identity element ($T(x_1, 1) = x_1$). All these properties which clearly agree with the boolean \wedge ones.

Some important *T-norm* examples (which are also implemented in SBR) are:

- Minimum *T-norm*: $T_{min}(x_1, x_2) = \min\{x_1, x_2\}$
- Product *T-norm*: $T_{prod}(x_1, x_2) = x_1 \cdot x_2$
- Lukasiewicz *T-norm*: $T_{Luk}(x_1, x_2) = \max\{0, x_1 + x_2 - 1\}$

For example the Product *T-norm*, used in most of the experiments will behave like this:

$$\begin{aligned}
\neg x_1 & \xrightarrow{\text{mapped}} N(x_1) = 1 - x_1 \\
x_1 \wedge x_2 & \xrightarrow{\text{mapped}} T_{prod}(x_1, x_2) = x_1 \cdot x_2 \\
x_1 \vee x_2 \equiv \neg(\neg x_1 \wedge \neg x_2) & \xrightarrow{\text{mapped}} N(T_{prod}(N(x_1), N(x_2))) = \quad (3.19) \\
& 1 - ((1 - x_1)(1 - x_2)) = \\
& x_1 + x_2 - x_1 \cdot x_2
\end{aligned}$$

Let us consider the logical implication $x_1 \Rightarrow x_2$. The natural approach would be to rewrite this formula as $\neg x_1 \vee x_2$ and treat as seen above. This will end up in

$$\neg x_1 \vee x_2 \xrightarrow{\text{mapped}} 1 + x_1 \cdot x_2 - x_1 \quad (3.20)$$

which corresponds to the implication but does not completely capture the inference process performed in a probabilistic or fuzzy logic context. There exists indeed a binary function for each T -norm that plays the role of implication, the *residuum*. Given two logical sub-formulas x_1 and x_2 in fuzzy logic the *residuum* is defined as:

$$\begin{cases} 1 & x_1 \leq x_2 \\ R & x_1 > x_2 \end{cases} \quad (3.21)$$

where R is specific for each T -norm. In particular:

- Minimum T -norm Residuum: $R = x_2$
- Product T -norm Residuum: $R = \frac{x_2}{x_1}$
- Lukasiewicz T -norm Residuum: $R = 1 - x_1 + x_2$

In the experimental section we used the Product T -norm Residuum when we need to translate an implication in constraints.

It is interesting to notice that, being the constraints applied on unsupervised examples, once the variable are grounded, the values used for converting the constraints into real values are the ones coming from the decision function associated to the predicate². Recovering the example used in Equation 3.17 the quantifiers-free part, using Product T -norm Residuum,

²In SBR is also possible to mix learned information (values come from the decision functions) or given information. In this case the values used come from the labeling of the data

can be mapped into:

$$\begin{cases} 1 & f_A(x_1) \cdot f_A(x_2) \leq f_B(x_1) \cdot f_B(x_2) \\ \frac{f_A(x_1) \cdot f_A(x_2)}{f_B(x_1) \cdot f_B(x_2)} & otherwise \end{cases} \quad (3.22)$$

where f_A, f_B are the decision functions for the predicate A and B .

Quantifiers in SBR

In addition to logical connectives results coming from *fuzzy logic* and T -norms, SBR implements a mapping also for the quantifiers that are needed for expressing properties on the whole dataset.

The universal quantifier in SBR is a measure of the violation of the constraints on the dataset. Let $\forall v.E(v, \mathcal{P})$ be an universal quantified formula corresponding to the h -th constraints. Here E is a logical expression over the variable v and the predicates \mathcal{P} , which, once v is grounded, corresponds to a *Fuzzy logic* formula and can therefore be mapped into a real value by the function $f_E(\mathbf{x}, \mathcal{P})$, where \mathbf{x} represents the grounding of v s.t. $\mathbf{x} \in \mathcal{S}$. The mapping can be now expressed as:

$$\forall v.E(v, \mathcal{P}) \xrightarrow{\text{mapped}} \phi_h(\mathbf{f}, \mathcal{S}) = \sum_{\mathbf{x} \in \mathcal{S}} 1 - t_E(\mathbf{f}, \mathbf{x}) \quad (3.23)$$

In the general case where multiple universal quantifier can be expressed, summation are nested for each quantifier. In particular, the mapping for $\forall v_1 \dots v_n.E(v_1 \dots v_n, \mathcal{P})$ will be:

$$\phi(\mathbf{f}, \mathcal{S}) = \sum_{\mathbf{x}_1 \in \mathcal{S}_1} \dots \sum_{\mathbf{x}_n \in \mathcal{S}_n} 1 - t_E(\mathbf{f}, \mathbf{x}_1 \dots \mathbf{x}_n) \quad (3.24)$$

Let us take for example the Equation 3.17, the corresponding mapping will

be:

$$\phi(\mathbf{f}, \mathcal{S}) = \sum_{\mathbf{x}_1 \in \mathcal{S}_1} \sum_{\mathbf{x}_2 \in \mathcal{S}_2} \begin{cases} 0 & f_A(x_1) \cdot f_A(x_2) \leq f_B(x_1) \cdot f_B(x_2) \\ 1 - \frac{f_A(x_1) \cdot f_A(x_2)}{f_B(x_1) \cdot f_B(x_2)} & otherwise \end{cases} \quad (3.25)$$

The existential quantifier express the presence of a variable that verifies the statement. In a *Fuzzy logic* contest, the intuition is that the desired result would be to return the value of the most satisfied statement. Indeed, we can formulate $\exists v.E(v, \mathcal{P})$ as:

$$\phi(\mathbf{f}, \mathcal{S}) = \min_{\mathbf{x} \in \mathcal{S}} 1 - t_E(\mathbf{f}, \mathbf{x}) \quad (3.26)$$

that can be in a very natural way extended to the n-existential quantifier \exists_n as:

$$\phi(\mathbf{f}, \mathcal{S}) = \sum_{\mathbf{x} \in \mathcal{S}_{n \min}} 1 - t_E(\mathbf{f}, \mathbf{x}) \quad (3.27)$$

where $\mathcal{S}_{n \min} = \arg \min_n \sum_{\mathbf{x} \in \mathcal{S}} 1 - t_E(\mathbf{f}, \mathbf{x})$, i.e. the first n assignment of \mathbf{x} in the dataset that more satisfy the constraint. It is interesting to notice that taking the two extreme values for n i.e. $n = 1$ and $n = |\mathcal{S}|$ we obtain respectively the existential quantifier and the universal one.

Chapter 4

Methods

4.1 Problem definition

As mentioned in the introduction, protein function prediction is a central but yet hard task in bioinformatics. There are plenty of papers about this topic and each has a slightly different interpretation of the problem. We have now given the necessary tools for understanding the definition of our problem.

Definition 4.1 (Gene Ontology cut) *Consider the labeled dataset $D_{\mathcal{L}} = \{(p_1, S_1), (p_2, S_2), \dots\}$ where p_i is the i -th protein and S_i the set of its annotation. Then, given a set of namespaces \mathcal{N} , a level threshold l and a count threshold c , then we call \hat{GO} a subgraph of GO such that:*

$$\hat{GO}(D, \mathcal{N}, l, c) = \{t \mid t \in GO \wedge level(t) \leq l \wedge |proteins(D, t)| \geq c\} \quad (4.1)$$

where $level$ and $proteins$ are functions that return respectively the level and the proteins associated with a Gene Ontology term.

This limitation is necessary for practical reasons. A minimum number of examples of each class in each fold is needed for a correct training and the level limit prevents to analyze too many predicates that would make the learning phase computationally too heavy.

The set of annotation S_i must be consistent, that means that if the term $t \in \hat{GO}(\mathcal{N}, l, c)$ belongs to S_i then all its ancestors must be in S_i . This consistency property is inherited from the True Path Rule of Gene Ontology explained before. Proteins can be described with a set of ‘fair’ features, i.e. features that it is reasonable to believe that unseen proteins will be in possession of.

It is now possible to define our problem, both from a conceptual and from a more practical point of view.

Definition 4.2 (Single Protein Function Prediction) *Given the labeled dataset $D_{\mathcal{L}}$, a GO cut $\hat{GO}(\mathcal{N}, l, c)$ and a previously unseen protein p with its features, the Single Protein Function Prediction (SPFP) problem consists in finding the consistent subgraph $\hat{S} \subseteq \hat{GO}(D, \mathcal{N}, l, c)$ that most likely represents the set of the annotations of p .*

This is a typical definition of an inductive classification task, that clearly express the concepts aiming our work. However, SBR provides additional tools that we have exploited to improve the quality of the prediction. The problem can be therefore reformulated by adding the technical details as follows.

Definition 4.3 (Multiple Protein Function Prediction) *Consider a labeled dataset $D_{\mathcal{L}}$ and a GO cut $\hat{GO}(\mathcal{N}, l, c)$. Now, given a set of unseen proteins $D_{\mathcal{U}}$ and a set of constraints Φ , the Multiple Protein Function Prediction (MPFP) problem consists in finding, for each $p \in D_{\mathcal{U}}$, the subgraph $\hat{S} \subseteq \hat{GO}(D, \mathcal{N}, l, c)$ consistent with the constraints Φ , that most likely represents the set of the annotations of p .*

The multiple protein set up allows us to use SBR in transductive learning mode. Moreover, as we have seen in Section 3.2, the unlabeled examples are fundamental to effectively impose constraints into the learning process.

Domains

Domain-based kernel are known in the literature for well performing in Molecular Function prediction. Ours, in particular, is based on the data that can be retrieved from InterPro [21] when an amino acid sequence is submitted.

InterPro offers a web service where amino acids sequence can be submitted and, via the *HMMER3* [10] algorithm, compared to the domain, family and superfamily data contained in its database. The results are exported in *xml* format and for each protein are associated the corresponding information retrieved. Once all proteins p_i have been analyzed, the Gram matrix for this kernel can be defined as:

$$G_{i,j} = \frac{|A_i \cap A_j|}{\sqrt{|A_i|^2 \cdot |A_j|^2}} \quad (4.2)$$

where A_k is the set of annotation retrieved in InterPro for the k -th protein.

Spectrum

Spectrum kernel rely only on the amino acid sequence of the protein as described in Section 3.1. In our experiment we used the normalized version of the string kernel function (no mismatch) shown in Section 3.1:

$$k^*(p_i, p_j) = \frac{k(p_i, p_j)}{\sqrt{k(p_i, p_i)k(p_j, p_j)}} \quad (4.3)$$

Being probably the simplest sequence-based we as a baseline for this kind of methods.

Protein complexes

The protein-complexes kernel is a diffusion kernel [24] on the protein complexes graph of the yeast [34]. As can be notice from Figure 2 (Appendix) this kernel is the most sparse, it is almost punctiform. This is due to the fact that protein complexes do not form a dense graph, it is indeed better described by the concept of graph forest, i.e. a set of (small) graphs

not interconnected. The biological explanation of this behavior is that, despite not being fixed, protein complexes subunits are usually part of just one complex and not exchanged between different complexes.

Microarray

The Microarray kernel matrix has been constructed computing the correlation for the gene expression data between each pair of proteins. Microarray data has been extracted from [43, 12] and covers the cell-cycle regulation and the response to environmental changes.

Let X and Y be the gene expression vectors for the protein p_i and p_j then the kernel function is:

$$k(p_i, p_j) = \frac{1}{n} \sum_{x \in X} \sum_{y \in Y} (x - \mu_X)(y - \mu_Y) \quad (4.4)$$

where n is the number of expression data per protein and μ_X the sample mean of X .

4.3 SBR rules

One of the fundamental contribute of this work is the integration of prior knowledge with the protein feature prediction process. Thanks to the Semantic-Based Regularization framework, we have been able to translate biological constraints coming from Gene Ontology and/or from Protein-Protein interaction network and incorporate them into the usual kernel methods machinery. In what follows are shown and explained the logical formulas that we used in our experiments.

4.3.1 Ontology Consistency

Gene Ontology is a trans-organism hierarchical ontology whose annotations are manually curated. These facts make GO a huge pool of information. Indeed we started exploiting the structure of the GO DAGs in order to force the consistency of the prediction, i.e. if an annotation α is predicted

for a protein p , then p should be predicted also in all the ancestors of α . Let us take for example the sub-graph of the cellular component DAG of Gene Ontology shown in Figure 4.2. If a protein p is predicted to be part of an *intra-cellular organelle*, by extension p is also part of all the GO terms bounded to the intra-cellular organelle one by an *is a* relation, that are *organelle* and *intra-cellular part*. Moreover, assuming that a protein has at least one path from the root to one of the leaves, if a protein p is predicted in a GO annotation α we can face two situations: 1) α is a leaf, base case, or 2) α is an inner node and that means that p should be in at least one of the children of α . These two constraints are strictly related to the structure of Gene Ontology and try to guarantee that the consistency of the predictions, from which the name of *Ontology Consistency rules*.

Let $Prot$ be the set of proteins, \mathcal{P} be the set of analyzed biological predicates (e.g. *catalytic activity*, *metabolic process*, ...), $chil : \mathcal{P} \rightarrow 2^{\mathcal{P}}$ a function that given a predicate return its children and $par : \mathcal{P} \rightarrow 2^{\mathcal{P}}$ a function that return the parents of a predicate. Given a predicate $u \in \mathcal{P}$,

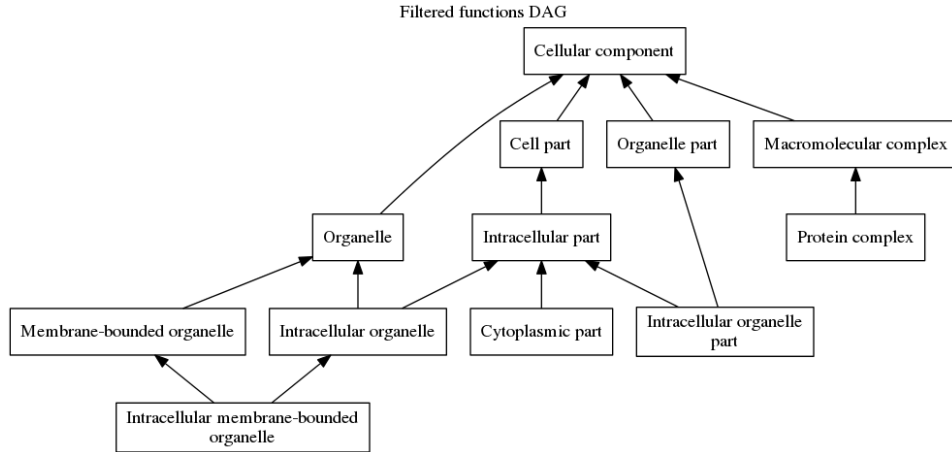


Figure 4.2: **Gene Ontology sub-graph.**

the first constrain can be expressed as:

$$\forall x \in Prot \bigwedge_{p \in par(u)} u(x) \Rightarrow p(x) \quad (4.5)$$

and the second as:

$$\forall x \in Prot \ u(x) \Rightarrow \bigvee_{c \in chil(u)} c(x) \quad (4.6)$$

The translation into a SBR-readable logic formula is fairly straightforward.

Formula 4.5 is rewritten as a list of single constraints in the form of

$$\forall x \in Prot \ Q(x) \Rightarrow P(x)$$

where P and Q are predicates s.t. $P \in par(Q)$, for each parent of the predicate Q in order to avoid the conversion in the conjunctive normal form.

Formula 4.6 remain substantially identical.

Bin nodes

For practical reasons working on the whole GO DAGs is unfeasible. In fact a minimum number of protein per annotation is necessary for producing reasonable predictions. Therefore in our experiment the analyzed predicate have always been a subset of the GO annotations. This pruning however produces an undesirable effect that makes Formula 4.6 no more valid. Indeed it is no more true that a protein associated with an inner-node of the GO DAG is at least associated to one of his children. It could be in fact the case that the pruned node was the only one associated with that protein.

In order to fix this drawback we introduce a class of nodes, whose goal is to gather all the proteins of the removed predicates, the *bin nodes*. Given a specific cut of the GO DAGs, the criteria for adding a *bin node* to annotation α are that α has at least one children node that survived to the cut and at least one who don't.

Let us expand the definitions of *par* and *chil* in such a way that $par : \mathcal{P} \rightarrow 2^{\mathcal{P} \cup \mathcal{B}}$ and $chil : \mathcal{P} \cup \mathcal{B} \rightarrow 2^{\mathcal{P}}$ where \mathcal{B} is the set of *bin nodes*. With this relatively small hack we are able to prune the GO DAGs without losing the consistency with the annotated dataset.

Trans-hierarchy relations

Since the first experiments, it was clear that some kernel performed way better on some GO DAG rather than an other. It is, for example, the case of the domain-based kernel whose performance on the molecular function DAG are considerably better when compared with the one obtained on the biological process DAG. We investigate therefore a way to propagate the information between the hierarchies. Gene Ontology structure came in our help providing a small but promising set of relations that could occur even between the GO DAGs. In particular these are *occurs in*, *part of* and *regulates*.

- The *occurs in* relation is a relationship that binds the biological process and the cellular component DAGs and therefore is not really useful in our experiment. Moreover there are just few relationships of this kind in the Gene Ontology.
- The *regulates* relation is relatively abundant and occurs between the hierarchies analyzed in this work. Unfortunately biological regulation takes place in a multitude ways that makes the translation of this relation into logical formula almost impossible.
- The *part of* relation instead is a good candidate for the trans hierarchy-rules, i.e. it occurs between molecular function and biological process DAGs with the meaning that a molecular function is part of some biological process.

The *part of* relation has therefore translated into the implication

$$\forall x \in Prot \ Q(x) \Rightarrow P(x)$$

for all predicate P, Q such that $P \in part_of(Q)$.

4.3.2 Protein-Protein Interaction

In order to expand the knowledge base we tried to integrate the Protein-Protein Interaction (PPI) data into the learning process. This information is partially already present into the complex kernel, but for the other kernels we decided to inject it in form of SBR constraints.

The motivations behind the idea of using this information can be found in the PPI network methods (Section 2.4). The underlying concept is that interacting proteins are likely to share their functions. From this idea we implemented the *PPI rules*. So, let BOUNDP be the binary predicate related to the interaction of proteins then:

$$\forall x, y \in Prot \ BOUND(x, y) \Rightarrow P(x) \Leftrightarrow P(y) \quad (4.7)$$

for all $P \in \mathcal{P}$. The translation in natural language of this constraint is that if two proteins interact then they share the predicate, for all the predicates.

Unfortunately this constraint is from a biological point of view not completely true. In fact it is true that interacting protein tends to share functions, but is usually not true that they share all of them. This results was also confirmed by the statistics gathered on the dataset. Table 1 and Table 2 show that, excluding the roots of the GO DAGs, there is no predicate that fully verifies the constraint 4.7. Moreover in Table 4.1 can be observed that Biological Process is more related to the interaction when compared to Molecular Function.

These results can be appreciated also from a biological point of view. Indeed, two interacting proteins very likely share a biological process, but it

	Average	Median	σ
Biological Process	0.629	0.615	0.274
Molecular Function	0.560	0.500	0.317

Table 4.1: **Jaccard coefficient on pairs of interacting proteins.** *The table shows the average, the median and the standard deviation of the Jaccard coefficient computed for each interacting pair of proteins in the dataset on the annotation set restricted to the analyzed GO DAG with cut 2,100*

is often not true that they are accomplishing the same molecular function. For example, muscle contraction is triggered by the arrival of a neurotransmitter in the neuromuscular junction, that connects the nervous systems and the muscular system via synapses. In the neuromuscular synapse there is the Nicotinic acetylcholine receptor, which is a protein complex composed by five subunits, arranged symmetrically around a central pore. Two of them, called α -subunits, are the only one able to bind the neurotransmitter (small molecule binding activity). The other three subunits have just structural roles. This is an example of proteins interacting in the same biological process, which does not share any molecular function.

According to the results obtained from the statistics gathered for the previous constraints we decided to weaken the constraints and reduce it to the following constraint:

$$\forall x, y \in Prot \text{ BOUND}(x, y) \Rightarrow \bigvee_{P \in \mathcal{P}|_{BP}} P(x) \cap P(y) \quad (4.8)$$

where $\mathcal{P}|_{BP}$ are the predicate examined with that are in the Biological Process DAG. We call this constraint DPP constraint.

Both the PP and the DPP constraint have been applied in combination with the OC rules and using given BOUND (true values), which are called PP1 and DPP1, or by learning them, PP2 and DPP2.

4.4 Measuring performances

For a robust assessment of the performance of our setup we used a 10-fold inner cross-validation procedure. The pseudo code of the fold generation algorithm is reported in Algorithm 2. In order to be consistent with the original task the final statistics are computed only over the investigated predicates, i.e. the *bin nodes* have been excluded being technical artifacts. The performance of the prediction of protein interaction are also taken as separated statistics.

Measuring performance in a multi-labeled setting is not trivial. There is indeed the new concept of *partially correct* that is not present in binary or multi-class problems [29]. The results can be analyzed from two different points of view, and from each of these we have different implementations of the common metrics used for assessing performances.

Some notation:

- n : the number of examples in the test set (we are using a cross validation and therefore $n = |D|$)
- k : the number of possible labels (the number of predicates analyzed in the experiment)
- Y_i : the set of labels of the i -th example (in our case the set of annotations for the i -th protein)
- Z_i : the set of predicted labels for the i -th example
- TP_j, TN_j, FP_j, FN_j : respectively the true positives, true negatives, false positives and false negatives of the j -th predicate.

Example based

Example based evaluation measures analyze the goodness of the results with respect to the examples. The extreme example of these class of metrics is the

Exact Match Ratio which return the ratio of completely correctly classified example with respect to the whole test set. Of course this measure too aggressive, especially for a hard task like this one.

We evaluated our results according to the metrics proposed by [13] that are:

$$\text{examples precision } P^\epsilon = \frac{1}{n} \sum_{i=1}^k \frac{|Y_i \cap Z_i|}{|Z_i|}$$

The *example precision* P^ϵ is the proportion of correctly predicted labels to the total of predicted labels, averaged over all examples.

$$\text{examples recall } R^\epsilon = \frac{1}{n} \sum_{i=1}^k \frac{|Y_i \cap Z_i|}{|Y_i|}$$

The *example recall* R^ϵ is the proportion of correctly predicted labels to the total of correct labels, averaged over all examples.

$$\text{examples } F1^\epsilon = \frac{1}{n} \sum_{i=1}^k \frac{2|Y_i \cap Z_i|}{|Y_i| + |Z_i|}$$

The *example F1* $F1^\epsilon$ is the harmonic mean of P^ϵ and R^ϵ .

An important statistics for this work is the measure of the consistency of the prediction for each example. It is indeed important to be able to quantify the improvements dictated by the introduction of constraints. Unfortunately the task is very specific to our problem and there are no used metrics, so we defined ourselves the consistency metrics as:

$$C^\epsilon = \frac{1}{n} \sum_{p \in Prot} \frac{1}{|\mathcal{P}(p)|} \sum_{n \in \mathcal{P}(p)} \begin{cases} 1 & \text{level}(n) = 1 \\ \frac{|parents(n) \cap \mathcal{P}(p)|}{|parents(n)|} & \text{otherwise} \end{cases} \quad (4.9)$$

where *parents* and *level* are respectively the functions returning the set of parents and the level in GO of a given predicate and \mathcal{P} the function returning the predicted predicates for a protein. This metric is completely

independent from the true values of the annotations, it just measures how much the consistency constraints are verified. The measure is unfortunately influenced by the level threshold used for the GO cut. It is indeed harder for deeper DAGs rather than for the ones with few levels to obtain high values for this metric. In the experimental phase it has been indeed used just to measure the changes inside a single experiment caused by rules.

Label based

Label based metrics focus themselves on the goodness of the predicate predictions. In this case we can apply the typical binary metrics and then compute the average between the predicates. There are two possible strategies for computing the average: 1) the *macro* average that simply computes the mean between the metrics for each predicate or 2) the *micro* average. In this case the confusion matrices for each predicate are summed up and on the obtained matrix are calculated the statistics. Here the definitions of the metrics used:

$$\begin{aligned}
P_\mu^\lambda &= \frac{\sum_{i=1}^k TP_i}{\sum_{i=1}^k TP_i + FP_i} & P_M^\lambda &= \frac{1}{k} \sum_{i=1}^k \frac{TP_i}{TP_i + FP_i} \\
R_\mu^\lambda &= \frac{\sum_{i=1}^k TP_i}{\sum_{i=1}^k TP_i + FN_i} & R_M^\lambda &= \frac{1}{k} \sum_{i=1}^k \frac{TP_i}{TP_i + FN_i} \\
F1_\mu^\lambda &= \frac{2 \sum_{i=1}^k TP_i}{\sum_{i=1}^k 2TP_i + FP_i + FN_i} & F1_M^\lambda &= \frac{1}{k} \sum_{i=1}^k \frac{2TP_i}{2TP_i + FP_i + FN_i}
\end{aligned}$$

There is no trivial choice of the best averaging method, so we have computed both.

In addition to the numerical metrics, we developed an algorithm for plotting macro averaged precision-recall curves. This kind of plot is very useful to understand the trend of the results, regardless to the decision threshold. Unfortunately there is no direct definition of the precision recall curve for multi-label problems, but [29] suggest that the definition of the

usual metrics used for binary problems can be extended to the multi-label case by averaging the results. We decide to use a *macro* average for this task because it comes more natural for computing the mean between curves.

The core idea of the algorithm we developed is to find, for each sampling coordinate on the recall axis, the best possible coordinate on the precision axis for each curve and then averaging them. So, let PRC_p the precision-recall curve for a predicate $p \in \mathcal{P}$ expressed a couple of list of points $PRC_p = ([P_1, \dots, P_n], [R_1, \dots, R_n])$, where (P_i, R_i) are respectively the the precision and recall coordinates of the i -th point int the list. Now suppose to have a set of PRC curves \mathcal{C} and n sampling points, then the average curve can be computed with the following algorithm:

Algorithm 1: *Macro* averaged Precision-Recall Curve algorithm

```

function Macro averaged Precision-Recall Curve (set  $\mathcal{C}$ , int  $n$ )
     $p \leftarrow [0]_n$  ; //  $(n+1)$ -dimensional array filled with zeros
     $r \leftarrow [i/n \mid i \in \{0, \dots, n\}]$  ; //  $(n+1)$ -dimensional array where
    the  $i$ -th position is  $i/n$ 
    foreach  $c_p, c_r \in \mathcal{C}$  do
        for  $k \in \{0, \dots, n\}$  do
             $i \leftarrow$  the index of the left nearest point in  $c_p$  to  $r[k]$ ;
             $j \leftarrow$  the index of the right nearest point in  $c_p$  to  $r[k]$ ;
             $f \leftarrow$ 
            function of the line passing through  $(c_p[i], c_r[i])$  and  $(c_p[j], c_r[j])$ ;

             $p[k] \leftarrow p[k] + f(r[k]) / |\mathcal{C}|$  ; // divides the contribute
            by the size of  $\mathcal{C}$ 
    return  $p, r$ 

```

In Figure 3 (Appendix) is shown how the algorithm performs when an increasing number of curves is averaged. It is interesting to notice how well the algorithm approximate the first curve, meaning that our approach is

performing reasonably good. Despite the variability of the curves, the macro averaged one can be still considered representative of the experiment.

4.5 Workflow description

During the development of the project, we implemented a pipeline for automatizing the execution of new experiments. The code has been written in python with the use of the following libraries: pygraphviz [16] (plotting of the statistic and result trees) and scikit-learn [32] (computing AUC metrics). In the following are described the most important steps of the developed pipeline.

Dataset selection and annotation

The dataset used in this work is a subset of the one used in [36], which consists in 1681 proteins of the *S. cerevisiae*. In order to be used in our experiments, the protein must have at least one non-electronically-inferred annotation in each of the analyzed GO domains that are Molecular Function and Biological Process. Unfortunately some of the 1681 initial proteins were not suitable for being used, therefore the dataset is dynamically adapted to fit the constraints. Despite this, the dataset is per-level consistent, i.e. experiments executed on the same depth level of the GO hierarchy use the same dataset.

Gene Ontology data structure

The obo-formatted files obtained from the Gene Ontology website has been initially parsed with *goatools*¹, an open-source python library created by Haibao Tang. During the implementation of the project this library has been widely modified to fit our needs.

Firstly, *goatools* takes into account only the *is a* relations between the GO term, but in our experiments we wanted to exploit different relations in order to expand the knowledge base. Therefore the data structure has

¹<https://github.com/tanghaibao/goatools>

Algorithm 2: Folds generation algorithm

```
function generate folds (int  $n$ , set  $\mathcal{Proteins}$ , list  $GO\ \mathcal{Terms}$ ,  
function  $f : \mathcal{P} \rightarrow 2^{\mathcal{T}}$ , function  $g : \mathcal{T} \rightarrow 2^{\mathcal{P}}$ )  
   $\mathcal{T} \leftarrow$  sorted  $\mathcal{T}$  according to  $|f(t)|$  for  $t \in \mathcal{T}$ ;  
  list  $folds \leftarrow [\{\}]_n$ ;  
  set  $selected \leftarrow \{\}$ ;  
  foreach  $t \in \mathcal{T}$  do  
     $to\ select \leftarrow g(t) \setminus selected$ ;  
    foreach  $p \in to\ select$  do  
      int  $index \leftarrow \underset{i}{\operatorname{argmin}}\{|folds[i]| \mid i \in 1, \dots, n\}$ ;  
       $folds[index] \leftarrow folds[index] \cup \{p\}$ ;  
       $selected \leftarrow selected \cup \{p\}$ ;  
  return  $folds$ 
```

been modified for keeping track of all the possible relations. After that, parametric accessors methods has been added to allow simple filtering of the related nodes. All the experiments of this work have been executed on a subset of the GO DAG hence a parametric pruning function has been added. This function takes as parameter the level and the count threshold and the GO namespaces to analyze. It returns a subset of the original GO DAG and for each GO term is associated the corresponding list of protein of the data set.

All these improvements proposal will be submitted to Haibao Tang at the end of the project.

Fold generation for cross validation

The SBR software does not provide an internal cross validation assessment, so it has to be done externally by splitting the data set in n folds and submitting to the software each time $n - 1$ folds as training set and one as test set.

Our experiments are intrinsically multi-label (each protein can be associated to more annotations), therefore producing n folds in where each class is balanced is very challenging. Trying to find a global optimum for all the

possible GO term -our classes- led to an intractable problem. So we decided to generate the folds experiment by experiment with a greedy approach that exploits the structure of the dataset. This led to a flexible and light system whose results well fits our goals.

The main idea behind the algorithm is to have a sufficient number of positive example of each class in each fold. The most critical classes are obviously the ones with less positive entries, which, in our setting, correspond to the deepest nodes in the GO hierarchy. Once the leaf examples are completely splitted into the folds, for the True Path Rule (discussed previously), the parent nodes are consequently splitted.

For these reasons the algorithm sorts the classes according to their positive example number in ascending order and starts splitting the first class. Then it continues iteratively until all the classes are separated in folds. The pseudocode in Algorithm 2 shows the details of the algorithm.

Execution

The training and testing of each fold is executed by the SBRS framework. The execution time grows fast with number of the rules, hence multiple SBRS executions are launched in parallel to soften this bottleneck. Each of the n execution produces a results file containing all the prediction values of the validation set that will be process afterwards.

Postprocessing

The final results are gathered and post processed according to the methodologies showed in methods. The statistics are stored in text format and plotted on the corresponding GO DAG for highlighting the dependencies between the nodes. An example is shown in Figure 4.3. At last experiments parameters and result metrics are stored in a local database.

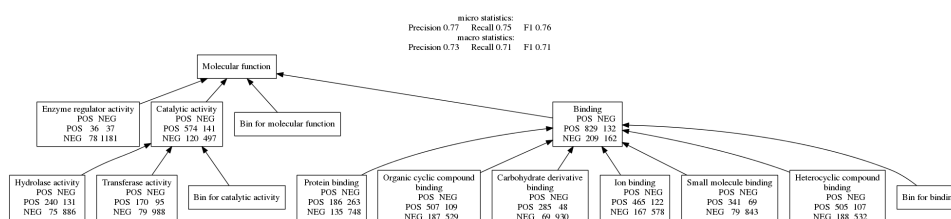


Figure 4.3: **Result tree example.** *The Central Dogma of Molecular Biology describes the three fundamental processes for life. Is interesting to notice that all these are carried out by proteins.*

Chapter 5

Results

The experimental phase of this work has been conducted on a subset (see Section 4.5 for the details) of the dataset used by [47], which consists of more than 1500 proteins of the *Saccharomyces cerevisiae* (baker yeast). The *S. cerevisiae* is a unicellular organism that belongs to the kingdom of *Fungi*. The baker yeast is a model organism for *eukaryotes*, it has the complexity of *eukaryotes* but grows easily in laboratories and has a short life cycle. Due to this reasons, *S. cerevisiae* was the first *eukaryotic* organism to be completely sequenced (in 1996 by [14]). Since then it has been widely studied and many specific resources can be find on the internet. Baker yeast represents therefore a perfect candidate for experimenting our setting.

As explained in Section 4.4, we performed a 10-fold cross validation for the assessment of the performance for the different rule sets and kernels in the analyzed Gene Ontology cuts. Unfortunately the computational time for each experiment is considerable high (in some cases it takes more then two hours for the training of each fold). Therefore we were not able to perform an exhaustive meta-parameter search to optimize the performances. We decided so to keep the default meta-parameter of SBR for each experiment, i.e. $\lambda_R = 1$ and $\lambda_C = 1$ (see Section 3.2).

The presentation of the results and the relative discussion will be divided

according to the two analyzed DAGs in Gene Ontology, Biological Process and Molecular Function. Only at the end we will show the results obtained combining them in a joint prediction task. The division comes from the fact that this hierarchies have deep differences from a biological point of view and therefore the tested kernels behave highly differently in the two circumstances. In each section, the discussion will be organized according to the following pattern.

- **Baseline results**

This portion of the discussion highlights the behavior of the kernel on the analyzed predicates in absence of rules. In this configuration SBR reduces to a l_2 regularized SVM task, indeed in absence of constraints, the learning problem becomes convex and converges to the global optimum. This analysis is fundamental for a complete understanding of the results obtained with the introduction of rules.

- **Ontology Consistency constraints**

The injection of Ontology Consistency (OC) constraints aims to improve the consistency of the predictions, with the final goal to refine the results through the propagation of information.

Consider the simple OC rule $h := \forall x. P_2(x) \Rightarrow P_1(x)$ where P_2 is a child node of P_1 in GO. Recall the definition of product residuum t-norm shown in Section 3.2, then the penalty function ϕ for h is:

$$\phi(h, \mathcal{S}) = 1 - \sum_{x \in \mathcal{S}} \begin{cases} 1 & P_2(x) \leq P_1(x) \\ \frac{P_1(x)}{P_2(x)} & otherwise \end{cases} \quad (5.1)$$

where \mathcal{S} is the set of unsupervised example on which the predictions for the predicates are computed. Consider now the case in which we have a strong positive prediction for the predicate P_2 and a slightly negative one for P_1 . We know, from the True Path Rule of Gene ontology, that if a term is predicted, then all its ancestors should be predicted too. Indeed the constraints h will contribute to the refinement of the results until $P_2(x) \leq P_1(x)$ will be satisfied. From this scenario the possible results on P_1 are:

- The positive prediction of $P_2(x)$ is correct (True Positive) and in the constrained gradient descent process the prediction of $P_1(x)$ is refined until the reaching of a value very close to $P_2(x)$ (low

contribute) or greater of $P_2(x)$ (no contribute). In both cases $P_1(x)$ changed its prediction to positive, hence a True Positive.

- The positive prediction of $P_2(x)$ is incorrect (False Positive) and in the constrained gradient descent process the prediction of $P_1(x)$ is forced to a positive classification value. In this case the action of the constraint badly impact on the performances. We end up indeed with two False positive, instead of just one. Nonetheless we are achieving a better consistency in the results.

The goal of this portion of the discussion is indeed to analyze the effects of this propagation of information on the predictions. As just explained, there are no guarantees that a better consistency will be reflected in a better quality of results.

- **Protein-Protein Interaction constraints**

The goal of these sets of rules is to increase the knowledge base injecting information on the PPI into the learning process. As explained in the Section 4.3 we applied these rules only on biological process to achieve biological consistency. In particular we used two supersets of constraints which differ by the base assumptions of the rules. In the first one indeed, the interacting proteins are supposed to share all biological processes. On the other hand, the second one assumes that two interacting proteins should be involved in at least one common biological process. The second one is more valid from a biological point of view, but less powerful. We will propose also the protein-protein interaction results to understand whether or not this predictions are influenced by the kernel used for the protein function prediction.

Notation summary

We will propose here a short summary of the notation used in this chapter to facilitate the reading and the comprehension of the results.

Rules (See Section 4.3)

- \mathcal{E} : no rules
- OC : Ontology Consistency rules
- PPI : Protein Protein Interaction rules
- DPPI : DNF Protein Protein Interaction rules

Metrics (See Section 4.4)

- Y^ϵ, Y^λ : respectively experiment- and label-based metrics
- Y_μ, Y_M : respectively micro- and macro-averaged metrics
- $P, R, F1$: respectively precision, recall and F1 measure
- \mathcal{C} : consistency measure

Gene Ontology (See Section 2.3)

- BP : Biological Process hierarchy
- MF : Molecular Function hierarchy
- $H_{l,c}$: hierarchy cut with level threshold l and count threshold c

5.1 Biological Process

In this section we will propose and discuss the results obtained with our system using as label space the Biological Process DAG of Gene Ontology. In particular we limited the set of analyzed annotation to the GO cut with count threshold $c = 100$ and level threshold $l \in \{2, 3\}$. We will focus on the impact of the different kernels and sets of rules in this situation.

The list of the analyzed predicate and the complete result tables can be found in the Appendix.

Baseline results

Level	Kernel	P_μ^λ	R_μ^λ	$F1_\mu^\lambda$	P_M^λ	R_M^λ	$F1_M^\lambda$
2	Complexes	0.460	0.932	0.616	0.435	0.920	0.557
	Domains	0.575	0.682	0.624	0.520	0.656	0.558
	Microarray	0.453	0.568	0.504	0.440	0.576	0.460
	Spectrum $k = 3$	0.569	0.536	0.552	0.496	0.458	0.457
3	Complexes	0.352	0.928	0.510	0.332	0.914	0.458
	Domains	0.493	0.705	0.58	0.442	0.681	0.517
	Microarray	0.361	0.575	0.443	0.341	0.565	0.396
	Spectrum $k = 3$	0.498	0.487	0.492	0.421	0.399	0.396

Table 5.1: **Baseline results for Biological Process.** *The table shows the result metrics for the experiments without rules in the Biological Process DAG.*

We can appreciate from Table 5.1 that the complex-based kernel and the domains one outperform the others in absence of constraints. Despite having similar $F1$ measures (especially within level 2) their behavior is very different. Indeed the complex kernel has an extremely high recall (more than 90%), but a very modest precision. Domains kernel instead behaves more uniformly.

The unbalancing in the complex kernel is explained by the fact that the information that it contains is very sparse (see heatmap in Appendix Figure

Level	Kernel	P_μ^λ	R_μ^λ	$F1_\mu^\lambda$	P_M^λ	R_M^λ	$F1_M^\lambda$
2	Complexes	0.795	0.839	0.816	0.722	0.804	0.754
	Domains	0.637	0.645	0.641	0.572	0.616	0.577
3	Complexes	0.734	0.833	0.781	0.656	0.79	0.709
	Domains	0.572	0.671	0.617	0.516	0.647	0.562

Table 5.2: **Filtered baseline results for Biological Process.** *This table shows the filtered version of the previously proposed ones. The undecided entries have been removed from the computation of the statistics, highlighting therefore the informative results*

1a). From the learning process point of view this is translated into the impossibility to take any decision for proteins whose kernel matrix row is zero in each entry, biologically translated, the proteins for which there are no interaction information. The effect on the results is that many protein predictions converge in the prior, which is also the threshold. As convention the threshold is considered positive and therefore these examples are considered positive predictions, hence the high recall and low precision. Differently, if the ‘undecided’ proteins are left apart, we end up with the results shown in Table 5.2. Here the situation changes drastically and the complex kernel shows its potential. Indeed, with a $F1_\mu$ around the 80% widely overtakes the the rivals. Figure 5.1 highlights this behavior. The Precision-Recall curve of the complexes kernel remains stable over the 70% of precision until reaching almost a recall of 80%.

Unfortunately we can not take this table as absolute value, because depending on how undecided values are treated we will definitively loose in precision or in recall, but it confirms the hypothesis that protein complexes are strongly related to biological processes. On the other hand it is interesting to notice that protein domains are, as one could expect, just marginally related to biological process. It is indeed true that proteins with similar domains may be part of completely different and not overlapping biological processes.

Level	Kernel	P^ϵ	R^ϵ	$F1^\epsilon$	$f P^\epsilon$	$f R^\epsilon$	$f F1^\epsilon$
2	Complexes	0.547	0.941	0.692	0.812	0.855	0.833
	Domains	0.604	0.688	0.643	0.631	0.652	0.641
	Microarray	0.469	0.578	0.518	0.469	0.578	0.518
	Spectrum $k = 3$	0.573	0.549	0.560	0.573	0.549	0.560
3	Complexes	0.467	0.940	0.625	0.759	0.855	0.804
	Domains	0.535	0.707	0.609	0.566	0.673	0.615
	Microarray	0.367	0.585	0.451	0.367	0.585	0.451
	Spectrum $k = 3$	0.503	0.500	0.501	0.503	0.500	0.501

Table 5.3: **Example-based metrics on baseline results.** *The prefixed f stays for filtered.*

Moreover, even though it has been shown that gene expression data are correlated to biological processes [44, 39, 22, 11], the microarray kernel performs poorly, even worse than the spectrum kernel. This may be partially explained by the combination of the noise that the corresponding heatmap highlights and the suboptimal parameter used in the experiments. A better choice of the λ_R meta-parameter could indeed partially soften the effect of the noise in the data.

Table 5.3 reports the example-based metrics for these experiments, with

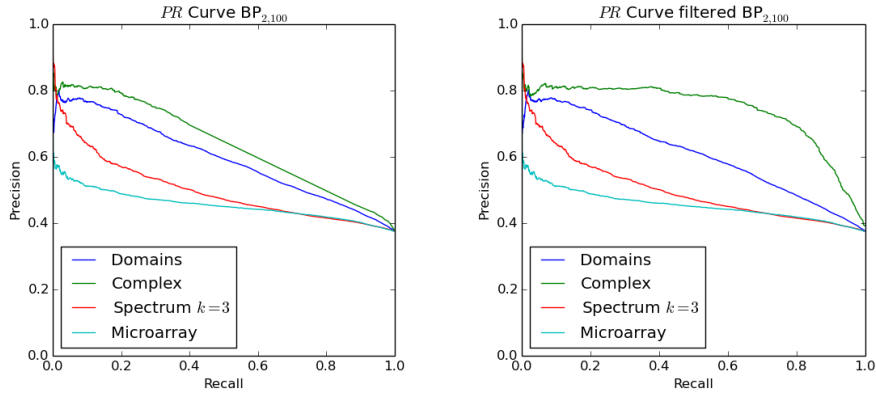


Figure 5.1: **Precision-Recall Curve of the kernel functions for Biological Process $_{2,100}$.** *In the two plots are reported the PR curves obtained on $BP_{2,100}$ with the empty set of rules. The right plot highlights the effect of the filtering on the curves*

and without the filtering of the undecided examples. The example-based metrics differ from the label-based one by the fact that the former measure the quality of the prediction per protein, the latter per predicate. The overall results (no filtering) highlight the differences in the behavior of the two typologies of metric. Being the average of the results obtained per single protein, they do not take into account the differences of the annotation set sizes.

Ontology Consistency constraints

	\mathcal{C}_ε	\mathcal{C}_{OC}	% increment
Complexes	0.2006	0.2006	0.000
Domains	0.2482	0.2537	2.198
Microarray	0.3204	0.4359	36.03
Spectrum $k = 3$	0.0815	0.1083	32.77

Table 5.4: **Consistency of the predictions with OC rules on BP_{2,100}**. The table shows the comparison of the consistency measure \mathcal{C} (see Section 4.4) in the experiments with and without consistency rules. The last column is the percentage increase obtained with the rules.

Table 5.4 proposes the values relative to the Ontology Constraints infringement metric \mathcal{C} . We can easily notice the difference in the impact of the rules on the kernel. Recall that in the experimental phase no meta-parameter tuning has been done and these results have been obtained therefore with the default meta-parameter of SBR. In particular, in this paragraph the λ_C parameter plays a crucial role giving more or less importance to the constraints, and a suboptimal choice deeply impacts on the results. We can indeed notice the different behaviors of the kernels. Complexes and domains kernel are just minimally affected by the constraints. In these kernel in fact, the information is very concentrated and precise. In both cases, the constraints meta-parameter is too soft to induce a substantial variation. On the other hand, in microarray and spectrum kernel information is much more diluted and the λ_C is sufficient to highly influence the results. As explained before,

Rules	Kernel	P_μ^λ	R_μ^λ	$F1_\mu^\lambda$	P_M^λ	R_M^λ	$F1_M^\lambda$
\mathcal{E}	Microarray	0.453	0.568	0.504	0.440	0.576	0.453
	Spectrum $k = 3$	0.569	0.536	0.552	0.496	0.458	0.569
OC	Microarray	0.475	0.621	0.538	0.439	0.594	0.475
	Spectrum $k = 3$	0.601	0.621	0.611	0.496	0.491	0.601

Table 5.5: **Effect of the constraints on Biological Process** _{2,100}.

the fact that the constraints are more respected do not imply that the results will benefit in terms of prediction performance. The hope is that the dominant part of the implication is correct and therefore also the information propagated.

Table 5.5 shows an extract of the comparison between the results obtained with and without OC rules. The complexes and domains kernels have been excluded, because unchanged from a practical point of view. The complete results can be found in Appendix. It is easy to notice that the OC constraints not only impacted positively on the consistency of the prediction, but also on the quality. We get a 12.5% improvement in the $F1_\mu^\lambda$ for the spectrum kernel and a 6.5% for the microarray one.

Protein-Protein Interaction constraints

Rules	P_μ^λ	R_μ^λ	$F1_\mu^\lambda$	P_M^λ	R_M^λ	$F1_M^\lambda$
\mathcal{E}	0.575	0.682	0.624	0.520	0.656	0.558
OC	0.575	0.682	0.624	0.521	0.656	0.559
PPI	0.573	0.679	0.621	0.519	0.655	0.556
DPPI	0.625	0.698	0.660	0.544	0.637	0.577

Table 5.6: **Effect of the OC rules on Biological Process** _{2,100}.

Table 5.6 reports an extract of the effects of the protein-protein interaction rules. We focus our attention on the domain kernel because it is the one best performing, but analogous consideration can be applied to all the others. Moreover the complete results can be found in the Appendix. It can

be noticed that the set of rules that obtained the highest $F1$ scores is the DPPI one. Unfortunately, these results have been obtained with a different version of SBR, which admits DNF constraints formulas and uses different t-norms, and therefore non completely comparable with the previous examples. Nonetheless we can appreciate the improvement compared to the PPI rules, which worsen the overall prediction. As reported in Table 4.1, the PPI constraints are often violated already in the training set. The scarce validity of the rule leads to a flow of misleading information that inevitably will worsen the prediction quality.

The prediction of protein-protein interaction has not been influenced by the underlying protein kernel. Indeed, we obtained the same results with all of them, meaning that we have been able to achieve a flow of information from the interactions to biological process but not vice versa. For the interaction prediction we obtained a $F1$ measure of circa 76%, both with the PPI constraints and with the DPPI.

5.2 Molecular Function

This section will propose the results for the Molecular Function DAG, we will focus especially on the GO cut for MF with count threshold $c = 100$ and level threshold $l \in \{2, 3, 4\}$. In these experiments we considered also the fourth level of gene ontology, because the MF DAG with the count threshold we used is smaller and it was therefore computationally easier to analyze deeper levels. As for Biological Process, the analysis will be centered around the effects of different kernel functions and sets of rules.

Baseline results

In Table 5.7 we can clearly notice how the domain-based kernel widely outperforms the competitors. It shows a good balance in precision and recall and achieves a $F1_\mu$ measure of almost 75%. The obtained results confirm the hypothesis that protein domains are strongly related to molecular function.

Level	Kernel	P_μ^λ	R_μ^λ	$F1_\mu^\lambda$	P_M^λ	R_M^λ	$F1_M^\lambda$
2	Complexes	0.415	0.884	0.565	0.404	0.869	0.527
	Domains	0.700	0.805	0.749	0.650	0.783	0.701
	Microarray	0.437	0.565	0.493	0.418	0.545	0.452
	Spectrum $k = 3$	0.549	0.798	0.650	0.491	0.677	0.555
3	Complexes	0.347	0.874	0.496	0.337	0.861	0.464
	Domains	0.665	0.829	0.738	0.625	0.827	0.704
	Microarray	0.363	0.571	0.444	0.353	0.561	0.412
	Spectrum $k = 3$	0.520	0.730	0.608	0.477	0.624	0.525
4	Complexes	0.305	0.869	0.452	0.297	0.861	0.423
	Domains	0.641	0.841	0.727	0.604	0.839	0.695
	Microarray	0.324	0.562	0.411	0.314	0.560	0.384
	Spectrum $k = 3$	0.496	0.696	0.579	0.463	0.595	0.504

Table 5.7: **Baseline results for Molecular Function.**

It is interesting to notice that the spectrum kernel, despite its simplicity, with a 65% of $F1_\mu$ at level 2 has reasonably good performances. The explanation of this behavior can be found in the consideration made for the domains kernel. It is indeed true that the spectrum kernel is capable to detect conserved regions of sequences and therefore mimic the behavior of the domain kernel. From this observation we can conclude that molecular functions are easier to predict from protein sequences rather than biological processes.

Complexes- and microarray-based kernel do not seem to carry much information, but we know from the biological process analysis that the complex kernel suffers from a convergence of results in the threshold which deeply alter the statistics. We applied also here the filtering on the results. The results are reported in Table 5.8.

Here we can further appreciate the performances of the domains-based kernel. Indeed, removing from the analysis the entries converged to the threshold, despite losing a little recall, the precision is greatly increased reaching more than 82% for all the analyzed thresholds. This is translated

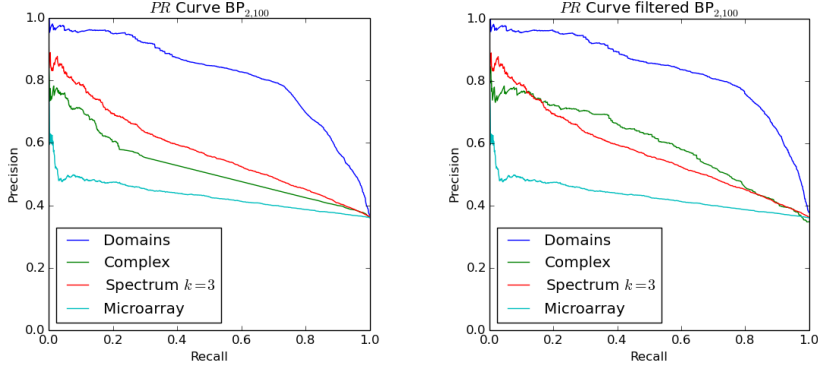


Figure 5.2: **Precision-Recall Curves of the kernel function for Molecular Function $_{2,100}$.** In the two plots are reported the PR curves obtained on $MF_{2,100}$ without rules. The right plot shows the effects of the filtering on the curves.

into an improvement of the $F1_\mu$ measure, which not only is stable over the 80%, but it increase with number of level analyzed. This particular is very interesting because it implies that the prediction of deeper nodes in the molecular function hierarchy are more related to domains than the higher ones.

Figure 5.2 shows the comparison of the precision-recall curves obtained with the different kernels with on without the filtering of the undecided values. The graphs confirms the observation made on the result tables. It is clear that in both conditions the domains kernel outperforms the others, maintaining a high precision (over 80%) until almost 80% o recall. Moreover

Level	Kernel	P_μ^λ	R_μ^λ	$F1_\mu^\lambda$	P_M^λ	R_M^λ	$F1_M^\lambda$
2	Complexes	0.622	0.676	0.648	0.568	0.635	0.590
	Domains	0.829	0.790	0.809	0.782	0.765	0.770
3	Complexes	0.566	0.650	0.605	0.510	0.606	0.548
	Domains	0.824	0.817	0.821	0.799	0.813	0.803
4	Complexes	0.531	0.639	0.580	0.482	0.600	0.530
	Domains	0.823	0.831	0.827	0.803	0.827	0.813

Table 5.8: **Filtered baseline results for Molecular Function.** This table reports the filtered version of the results presented in the previous one.

Rules	Kernel	P_μ^λ	R_μ^λ	$F1_\mu^\lambda$	P_M^λ	R_M^λ	$F1_M^\lambda$
\mathcal{E}	Complexes	0.415	0.884	0.565	0.404	0.869	0.527
	Domains	0.700	0.805	0.749	0.650	0.783	0.701
	Microarray	0.437	0.565	0.493	0.418	0.545	0.452
	Spectrum $k = 3$	0.549	0.798	0.650	0.491	0.677	0.555
OC	Complexes	0.415	0.885	0.565	0.404	0.870	0.527
	Domains	0.701	0.806	0.750	0.650	0.784	0.702
	Microarray	0.449	0.613	0.518	0.411	0.566	0.461
	Spectrum $k = 3$	0.568	0.769	0.653	0.503	0.645	0.552

Table 5.10: **Effects of the OC rules on Molecular Function** _{2,100}.

we can notice that, despite being just marginally related with the molecular function hierarchy, the complex kernel positively responds to the filtering also in this situation.

The analysis made on the results with our kernels confirms the trends obtained by most of the systems found in literature [6, 17], in which the prediction of molecular functions results to be an easier task if compared to the one of biological processes. A partial explanation this result lies into the consideration that molecular functions are intrinsic properties of the proteins and just marginally depend on the surrounding environment. Is therefore easier to detect important feature “in” the protein, which are also more abundant and precise thanks to the new sequencing techniques.

Ontology Consistency constraints

	$\mathcal{C}_\mathcal{E}$	\mathcal{C}_{OC}	% increment
Complexes	0.139	0.141	0.862
Domains	0.249	0.250	0.104
Microarray	0.266	0.407	52.96
Spectrum $k = 3$	0.081	0.170	110.8

Table 5.9: **Consistency of the predictions with OC rules on MF** _{2,100}. The table shows the comparison of the consistency measure \mathcal{C} (see Section 4.4) in the experiments with and without consistency rules. The last column is the percentage increase obtained with the rules.

Rules	λ_C	P_μ^λ	R_μ^λ	$F1_\mu^\lambda$	P_M^λ	R_M^λ	$F1_M^\lambda$	\mathcal{C}
\mathcal{E}		0.700	0.805	0.749	0.650	0.783	0.701	0.249
OC	10^1	0.698	0.828	0.757	0.646	0.798	0.706	0.358
	10^3	0.590	0.904	0.714	0.565	0.869	0.667	0.771
	10^5	0.444	0.968	0.609	0.423	0.929	0.566	0.995

Table 5.11: **Effects of λ_C on the OC rules.** *The table shows the effect of the constraint strictness on the prediction of $MF_{2,100}$ with the domains kernel.*

Table 5.9 and Table 5.10 report respectively the effect of the OC constraints on the overall consistency of the prediction and on their quality. Similar consideration to the ones made for the biological process case can be applied also here. This confirms the hypothesis that the right choice of the meta-parameter is more related to the kernel rather than to the analyzed task. We can indeed notice that complexes and domains kernel have been just marginally affected by the constraints, on the other hand on microarray and spectrum ones the OC rules deeply impacted on the consistency measure. However, in this case, the improvement in the consistency of the predictions is not completely transferred to their quality. The microarray kernel, which has the greatest gains in the consistency metric (in absolute value), has just a 2.5% increase in the $F1_\mu^\lambda$ measure.

The examples shown in Table 5.11 make the relation between the strictness of the constraints and the quality of the performance clearer. It is indeed easy to notice that the increase of the consistency measure do not directly implies a benefit in term of performance. The last row shows an experiment made on purpose to demonstrate that, despite being completely satisfied in the training set, hard constraints (the value of λ_C is so high to dominate in the optimization process) may strongly worsen the quality of prediction.

5.3 Joint prediction of Biological Process and Molecular Function

Part of the goals of this work was to analyze the effect of trans-hierarchy relationships on the final prediction of protein features. In this section we will therefore analyze the results obtained on the joint prediction of Molecular Function and Biological Process. Being the number of analyzed rules proportional to the number of predicates, the prediction of both hierarchies becomes soon computationally very hard. We decided therefore to limit the search to the first two levels of Gene Ontology with a count threshold $c = 100$. Unfortunately with this restrictions the only GO cross-hierarchy relation taken into account is:

$$\text{CATALYTIC-ACTIVITY}^{\text{MF}} \xrightarrow{\text{part of}} \text{METABOLIC-PROCESS}^{\text{BP}}$$

This scarceness, how will be shown, make the information flow between BP and MF just marginally effective.

In Table 5.10 are reported the results obtained in these experiments with and without constraints. It can be notice that the kernels reflect the observation made for the disjointed predictions. It is important to notice that the number of predicted predicates is larger in the Biological Process hierarchy making the contribute of Molecular Function slightly less significant.

The domains kernel obtains the bests $F1$ scores, both in the constrained and unconstrained cases, confirming its leading position in the unfiltered condition. Moreover, we can appreciate the effect of OC rules on the spectrum kernel, which gains almost a 5% in the constrained case. This confirm the results highlighted in the previous sections, i.e. the kernel most affected by constraints are smother ones.

The introduction of the DPPI rules proposes very interesting results.

Rules	Kernel	P_μ^λ	R_μ^λ	$F1_\mu^\lambda$	P_M^λ	R_M^λ	$F1_M^\lambda$
\mathcal{E}	Complexes	0.452	0.922	0.606	0.431	0.906	0.555
	Domains	0.622	0.724	0.669	0.571	0.702	0.613
	Microarray	0.458	0.568	0.507	0.446	0.571	0.467
	Spectrum $k = 3$	0.568	0.616	0.591	0.499	0.530	0.496
OC	Complexes	0.452	0.922	0.606	0.431	0.906	0.555
	Domains	0.622	0.724	0.669	0.571	0.701	0.613
	Microarray	0.475	0.608	0.533	0.444	0.584	0.479
	Spectrum $k = 3$	0.595	0.662	0.627	0.505	0.541	0.510
DPPI	Complexes	0.453	0.922	0.607	0.432	0.905	0.555
	Domains	0.667	0.750	0.706	0.592	0.694	0.633
	Microarray	0.467	0.570	0.513	0.440	0.551	0.464
	Spectrum $k = 3$	0.634	0.595	0.614	0.525	0.463	0.481

Table 5.12: **Results for native joint prediction of BP_{2,100} and MF_{2,100}**

Besides confirming that valid constraints positively impact on the overall quality of predictions, it shows that rules and kernels can effectively work together and mix their strengths. This behavior is evident when compared the two most informative kernels, i.e the complexes and the domains ones, in the DPPI setting. It can be noticed that the complexes kernel is just minimally affected by this set of rules, on the other hand the domains one has a very positive improvement. We know that the complexes kernel is based on the protein complexes interaction network that is a subset of the PPI. Therefore the increment of information in this case is marginal. Differently the domains kernel has no internal information on protein protein interaction. Hence the constraints inject a considerable quantity of new knowledge that is translated into an improvement of the overall predictions quality.

Chapter 6

Conclusion

The aim of this work was to exploit a state-of-the-art statistical relational learning framework to predict protein features. We focused our analysis on the prediction of the annotations belonging to the first levels of the biological process and molecular functions hierarchies of Gene Ontology, which is a structured vocabulary of the definition of genes products (in our case proteins). The framework employed in this work is called Semantic Based Regularization and allowed us to define first order logic rules that can constrain the typical kernel machine learning process.

The main efforts in this work have been oriented to the designing of effective but valid constraints, that could exploits different biological sources to refine the quality of prediction, and to the implementation of four kernel functions that covers various aspects of proteins nature. The whole setup has been then tested on more then 1500 proteins of the *S. cerevisiae*, the baker yeast.

From the experimental results we were able to appreciate the effects of the constraints on the predictions. Despite the suboptimal choice of the meta parameters, we noticed good improvements with the ontology consistency rules, in particular when applied in combination with ‘smoother’ kernel. Differently, with kernels where information is very sharp and defines, like the

domains and complexes ones, are needed higher values of the λ_C parameter to impact on the results. However, too strict constraints will considerably worsen the prediction. Moreover is important to verify the validity of the rules, because the injection of misleading information will clearly lead to a negative impact on the performance.

The prediction of the biological processes related to a protein confirms the trend found in the literature and was a complex task. Indeed the final prediction with the best kernel (domains kernel) were just over 60% of the $F1_\mu$ measure. However the strong unbalance of the complex kernel results led us to notice that for several proteins the system was not able to take a decision and the predictions converged in the threshold. This behavior can be attributed to the sparseness of this kernel and, being these proteins easily recognizable, they can be removed from the analysis. In this situation the complexes kernel shows its correlation to the biological processes and achieves an $F1_\mu$ measure around the 80%. The obtained results are fully admissible from a biological point of view. It is indeed true that domains are just marginally related to biological processes; two proteins can share very similar domains and, despite this, being part of two completely different biological process. On the other hand, two proteins part of the same protein complex are very likely to be part of the same biological process.

The prediction of molecular functions gave us good results. We have been able to predict them with more of the 70% of $F1_\mu$ measure in all the analyzed levels thanks to the domains kernel. This confirms the hypothesis that protein domains are highly related to their molecular function. It was interesting to notice that, on the filtered results, the quality of the results (always over the 80% of $F1_\mu$) increased with the depth of the levels. Our conclusion is that domains are so specific for the MF annotations, that high level nodes of Gene Ontology are too general and mislead some predictions. In addition to the domains kernel, it is interesting to notice the behavior of

the spectrum kernel. Indeed, despite its simplicity and the little information used (just the amino acids sequence), it obtains decent results. This leads to the conclusion that the amino acid sequence plays a crucial role in the prediction of molecular function, much more than for the biological process. It is indeed true that molecular functions strongly depends on the structure and composition of a protein and are just marginally influenced by the surrounding environment. Therefore the information needed to predict them can be found directly ‘in’ the protein, which thanks to the improvements in the sequencing techniques is always more abundant and precise. On the other hand biological processes strongly depends on the environment surrounding the protein. It is indeed the conjunction of subcellular localization, molecular functions and interactions, and the moment in the cell cycle to determine the biological processes a protein is part of. This make the task much more complex. Moreover the information are mostly located ‘outside’ the protein and therefore harder to retrieve.

Concluding, we obtained good results with the molecular function annotations, but more work has to be done in the prediction biological processes in order to achieve the reliability required to offer clear target that biologist could exploit to focus their laboratory researches. Nonetheless, this work shows the effectiveness of combining kernel functions and logical constraints in the prediction of biological features. Indeed, our results represent a good starting point for further works in this direction.

Appendices

Protein-Protein interaction

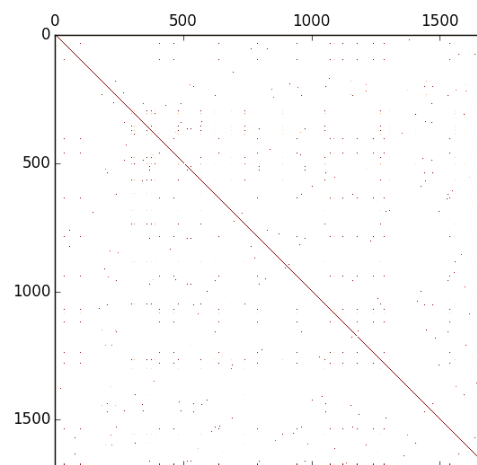
Level	Predicate	POS PP	TOT PP	RATIO
0	MOLECULAR-FUNCTION	2688	2688	1.000
1	BINDING	1848	2458	0.752
	CATALYTIC-ACTIVITY	693	1647	0.421
	ENZYME-REGULATOR-ACTIVITY	62	451	0.137
2	ORGANIC-CYCLIC-COMPOUND-BINDING	937	1815	0.516
	HETEROCYCLIC-COMPOUND-BINDING	936	1814	0.516
	ION-BINDING	657	1593	0.412
	PROTEIN-BINDING	448	1267	0.354
	TRANSFERASE-ACTIVITY	199	631	0.315
	SMALL-MOLECULE-BINDING	346	1212	0.285
	HYDROLASE-ACTIVITY	251	929	0.270
	CARBOHYDRATE-DERIVATIVE-BINDING	282	1061	0.266

Table 1: **PPI predicate sharing for Molecular Function_{2,100}**. *In this table are reported the values of the predicate coverage for the protein in the interaction dataset. The values has been calculated on the real predicate and proteins for the MF_{2,100}. TOP PP indicates the number of protein pair for which at least one of the two has the corresponding predicate. POS PP instead in the number of protein pairs having both the predicate. the last column is the ratio between the previous two.*

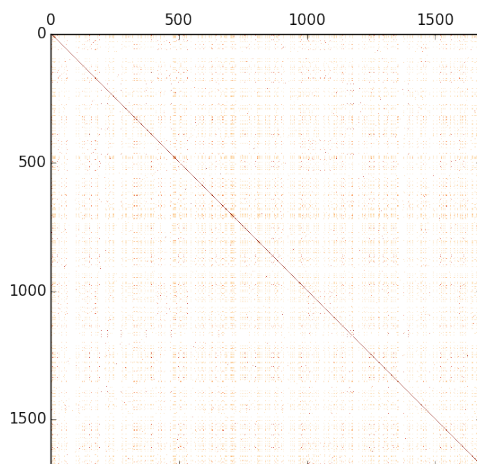
Level	Predicate	POS PP	TOT PP	RATIO
0	BIOLOGICAL-PROCESS	3125	3125	1.000
1	CELLULAR-PROCESS	2681	3036	0.883
	SINGLE-ORGANISM-PROCESS	1930	2615	0.738
	METABOLIC-PROCESS	1938	2661	0.728
	BIOLOGICAL-REGULATION	1233	2066	0.597
	LOCALIZATION	732	1354	0.541
	CELLULAR-COMPONENT-ORGANIZATION...	1162	2195	0.529
	RESPONSE-TO-STIMULUS	510	1312	0.389
2	CELLULAR-METABOLIC-PROCESS	1866	2607	0.716
	ORGANIC-SUBSTANCE-METABOLIC-PROCESS	1786	2517	0.710
	PRIMARY-METABOLIC-PROCESS	1738	2471	0.703
	SINGLE-ORGANISM-CELLULAR-PROCESS	1592	2437	0.653
	REGULATION-OF-BIOLOGICAL-PROCESS	1129	1963	0.575
	NITROGEN-COMPOUND-METABOLIC-PROCESS	1064	1882	0.565
	ESTABLISHMENT-OF-LOCALIZATION	654	1219	0.537
	CELLULAR-COMPONENT-ORGANIZATION	1105	2160	0.512
	BIOSYNTHETIC-PROCESS	533	1046	0.510
	SINGLE-ORGANISM-LOCALIZATION	540	1090	0.495
	SINGLE-ORGANISM-TRANSPORT	486	1010	0.481
	MACROMOLECULE-LOCALIZATION	425	895	0.475
	SINGLE-ORGANISM-METABOLIC-PROCESS	736	1657	0.444
	ORGANELLE-ORGANIZATION	623	1545	0.403
	CATABOLIC-PROCESS	496	1269	0.391
	CELLULAR-RESPONSE-TO-STIMULUS	436	1188	0.367
	RESPONSE-TO-STRESS	310	910	0.341
	CELLULAR-LOCALIZATION	233	689	0.338
	SINGLE-ORGANISM-MEMBRANE-ORGANIZATION	125	403	0.310
	REGULATION-OF-BIOLOGICAL-QUALITY	151	510	0.296
	RESPONSE-TO-CHEMICAL	78	387	0.202
	REGULATION-OF-MOLECULAR-FUNCTION	137	695	0.197

Table 2: **PPI predicate sharing for Biological Process_{2,100}**. *In this table are reported the values of the predicate coverage for the protein in the interaction dataset. The values has been calculated on the real predicate and proteins for the BP_{2,100}. TOP PP indicates the number of protein pair for which at least one of the two has the corresponding predicate. POS PP instead in the number of protein pairs having both the predicate. the last column is the ratio between the previous two.*

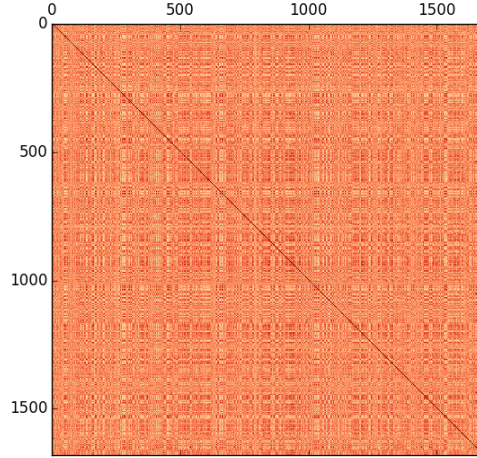
Kernel sparseness



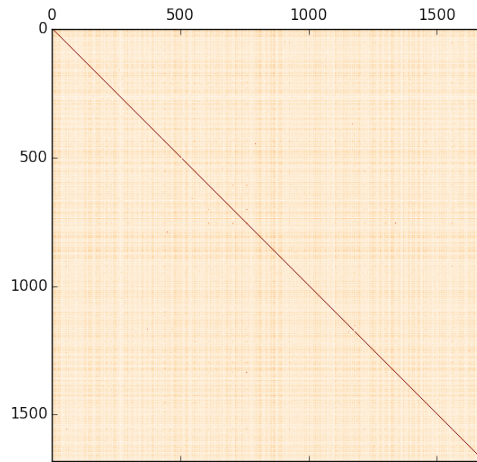
(a) Complex-based kernel



(b) Domain-based kernel



(a) Microarray kernel



(b) Spectrum kernel $k = 3$

Figure 2: **Kernel matrix heatmaps.** *The proposed heatmaps refer to the kernel matrix we implemented. The intensity of the color for each coordinate in the plots is proportional to the kernel value for that pair of points.*

Curve averaging algorithms

results

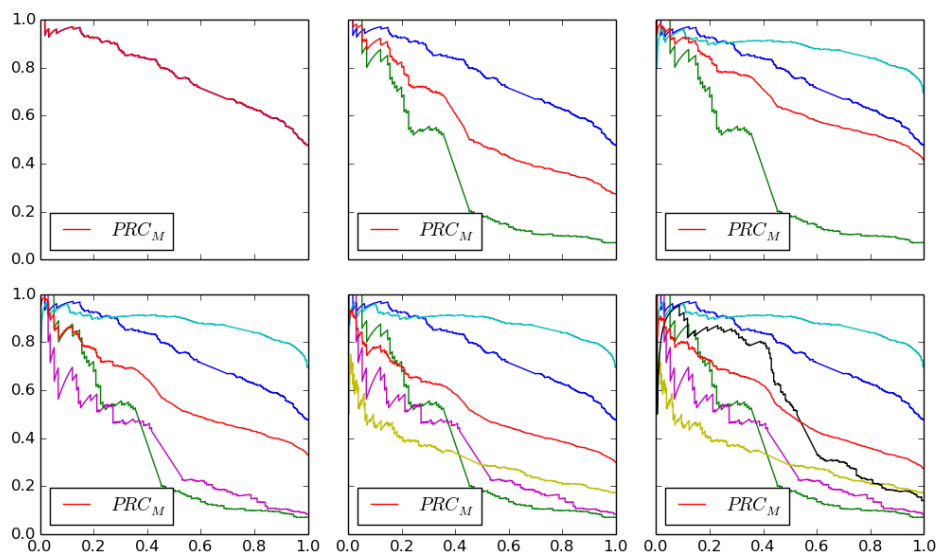


Figure 3: *Macro* averaged Precision-Recall Curve Example.

Results Tables

Kernel	Rules	Precision _{μ}	Recall _{μ}	F1 _{μ}	Precision _{M}	Recall _{M}	F1 _{M}
Complex	ϵ	0.415	0.884	0.565	0.404	0.869	0.527
	OC	0.415	0.885	0.565	0.404	0.870	0.527
Domains	ϵ	0.700	0.805	0.749	0.650	0.783	0.701
	OC	0.701	0.806	0.750	0.650	0.784	0.702
Microarray	ϵ	0.437	0.565	0.493	0.418	0.545	0.452
	OC	0.449	0.613	0.518	0.411	0.566	0.461
Spectrum $k = 3$	ϵ	0.549	0.798	0.650	0.491	0.677	0.555
	OC	0.568	0.769	0.653	0.503	0.645	0.552

Table 3: **Results Molecular Function_{2,100}**. Results obtained with the four kernels on the Molecular Function with `level_threshold = 2` and `count_threshold = 100`. The notation metric_μ stays for micro average and metric_M stays for macro average. The definition of the rules are reported in the Methods chapter (Section 4.3).

Kernel	Rules	Precision _{μ}	Recall _{μ}	F1 _{μ}	Precision _{M}	Recall _{M}	F1 _{M}
Complex	ϵ	0.347	0.874	0.496	0.337	0.861	0.464
	OC	0.347	0.874	0.496	0.337	0.862	0.464
Domains	ϵ	0.665	0.829	0.738	0.625	0.827	0.704
	OC	0.665	0.829	0.738	0.625	0.826	0.704
Microarray	ϵ	0.363	0.571	0.444	0.353	0.561	0.412
	OC	0.363	0.572	0.444	0.352	0.561	0.412
Spectrum $k = 3$	ϵ	0.520	0.730	0.608	0.477	0.624	0.525
	OC	0.520	0.730	0.607	0.478	0.623	0.525

Table 4: **Results Molecular Function_{3,100}**. Results obtained with the four kernels on the Molecular Function with `level_threshold = 3` and `count_threshold = 100`. The notation metric_μ stays for micro average and metric_M stays for macro average. The definition of the rules are reported in the Methods chapter (Section 4.3).

Kernel	Rules	Precision _{μ}	Recall _{μ}	F1 _{μ}	Precision _{M}	Recall _{M}	F1 _{M}
Complex	ϵ	0.305	0.869	0.452	0.297	0.861	0.423
	OC	0.305	0.869	0.452	0.297	0.861	0.423
Domains	ϵ	0.641	0.841	0.727	0.604	0.839	0.695
	OC	0.641	0.841	0.727	0.604	0.839	0.695
Microarray	ϵ	0.324	0.562	0.411	0.314	0.560	0.384
	OC	0.324	0.563	0.411	0.314	0.560	0.384
Spectrum $k = 3$	ϵ	0.496	0.696	0.579	0.463	0.595	0.504
	OC	0.497	0.696	0.579	0.463	0.593	0.504

Table 5: **Results Molecular Function_{4,100}**. Results obtained with the four kernels on the Molecular Function with `level_threshold = 4` and `count_threshold = 100`. The notation metric_μ stays for micro average and metric_M stays for macro average. The definition of the rules are reported in the Methods chapter (Section 4.3).

Kernel	Rules	Precision _m	Recall _m	F1 _m	Precision _M	Recall _M	F1 _M
Complex	ϵ	0.460	0.932	0.616	0.435	0.920	0.557
	OC	0.460	0.932	0.616	0.435	0.920	0.557
	PPI	0.460	0.932	0.616	0.435	0.920	0.557
	DPPI	0.461	0.934	0.618	0.436	0.918	0.558
Domains	ϵ	0.575	0.682	0.624	0.520	0.656	0.558
	OC	0.575	0.682	0.624	0.521	0.656	0.559
	PPI	0.575	0.682	0.624	0.521	0.656	0.559
	DPPI	0.625	0.699	0.66	0.544	0.638	0.578
Microarray	ϵ	0.453	0.568	0.504	0.440	0.576	0.460
	OC	0.475	0.621	0.538	0.439	0.594	0.476
	PPI	0.467	0.597	0.524	0.440	0.584	0.470
	DPPI	0.447	0.536	0.487	0.415	0.512	0.431
Spectrum $k = 3$	ϵ	0.569	0.536	0.552	0.496	0.458	0.457
	OC	0.601	0.621	0.611	0.496	0.491	0.482
	PPI	0.596	0.599	0.598	0.497	0.481	0.477
	DPPI	0.624	0.576	0.599	0.503	0.434	0.454

Table 6: **Results Biological Process_{2,100}**. *Results obtained with the four kernels on the Biological Process with level_threshold = 2 and count_threshold = 100. The notation metric_m stays for micro average and metric_M stays for macro average. The definition of the rules are reported in the Methods chapter (Section 4.3). The experiments conducted on the last sets of rules (DPPI) have been executed on a newer version of SBRS that supports constraints in DNF and uses a different t-norm. Therefore, the results of these experiments are non completely comparable with the others.*

Kernel	Rules	Precision _m	Recall _m	F1 _m	Precision _M	Recall _M	F1 _M
Complex	ϵ	0.352	0.928	0.510	0.332	0.914	0.458
	OC	0.352	0.928	0.510	0.332	0.914	0.458
	PPI	0.352	0.928	0.510	0.332	0.914	0.458
	DPPI	0.353	0.927	0.511	0.332	0.911	0.458
Domains	ϵ	0.493	0.705	0.580	0.442	0.681	0.517
	OC	0.493	0.704	0.580	0.442	0.681	0.517
	PPI	0.493	0.704	0.580	0.442	0.681	0.517
	DPPI	0.543	0.685	0.606	0.466	0.627	0.525
Microarray	ϵ	0.361	0.575	0.443	0.341	0.565	0.396
	OC	0.364	0.582	0.448	0.341	0.567	0.398
	PPI	0.362	0.578	0.445	0.341	0.566	0.397
	DPPI	0.356	0.582	0.442	0.328	0.554	0.387
Spectrum $k = 3$	ϵ	0.498	0.487	0.492	0.421	0.399	0.396
	OC	0.500	0.488	0.494	0.422	0.397	0.395
	PPI	0.500	0.487	0.494	0.423	0.396	0.395
	DPPI	0.560	0.479	0.516	0.443	0.341	0.373

Table 7: **Results Biological Process_{3,100}**. *Results obtained with the four kernels on the Biological Process with level_threshold = 3 and count_threshold = 100. The notation metric_μ stays for micro average and metric_M stays for macro average. The definition of the rules are reported in the Methods chapter (Section 4.3). The experiments conducted on the last sets of rules (DPPI) have been executed on a newer version of SBRS that supports constraints in DNF and uses a different t-norm. Therefore, the results of these experiments are non completely comparable with the others.*

Kernel	Rules	Precision _{μ}	Recall _{μ}	F1 _{μ}	Precision _{M}	Recall _{M}	F1 _{M}
Complex	ϵ	0.452	0.922	0.606	0.431	0.906	0.555
	OC	0.452	0.922	0.606	0.431	0.906	0.555
Domains	ϵ	0.622	0.724	0.669	0.571	0.702	0.613
	OC	0.622	0.724	0.669	0.571	0.701	0.613
Microarray	ϵ	0.458	0.568	0.507	0.446	0.571	0.467
	OC	0.475	0.608	0.533	0.444	0.584	0.479
Spectrum $k = 3$	ϵ	0.568	0.616	0.591	0.499	0.53	0.496
	OC	0.595	0.662	0.627	0.505	0.541	0.51

Table 8: **Results joint prediction of Biological Process and Molecular Function_{4,100}**. Results obtained with the four kernels on joint prediction of Biological Process and Molecular Function with *level_threshold* = 4 and *count_threshold* = 100. The notation *metric _{μ}* stays for micro average and *metric _{M}* stays for macro average. The definition of the rules are reported in the Methods chapter (Section 4.3).

Bibliography

- [1] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.
- [2] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.
- [3] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, TN Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.
- [4] UniProt Consortium et al. The universal protein resource (uniprot). *Nucleic acids research*, 36(suppl 1):D190–D195, 2008.
- [5] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
- [6] Domenico Cozzetto, Daniel WA Buchan, Kevin Bryson, and David T Jones. Protein function prediction by massive integration of evolutionary analyses and multiple data sources. *BMC bioinformatics*, 14(Suppl 3):S1, 2013.

- [7] Margaret O Dayhoff and Robert M Schwartz. A model of evolutionary change in proteins. In *In Atlas of protein sequence and structure*. Citeseer, 1978.
- [8] Michelangelo Diligenti, Marco Gori, Marco Maggini, and Leonardo Rigutini. Bridging logic and kernel machines. *Machine learning*, 86(1):57–88, 2012.
- [9] Francesco Dinuzzo and Bernhard Schölkopf. The representer theorem for hilbert spaces: a necessary and sufficient condition. In *Advances in neural information processing systems*, pages 189–196, 2012.
- [10] Sean R Eddy et al. A new generation of homology search tools based on probabilistic inference. In *Genome Inform*, volume 23, pages 205–211. World Scientific, 2009.
- [11] Michael B Eisen, Paul T Spellman, Patrick O Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.
- [12] Audrey P Gasch, Paul T Spellman, Camilla M Kao, Orna Carmel-Harel, Michael B Eisen, Gisela Storz, David Botstein, and Patrick O Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular biology of the cell*, 11(12):4241–4257, 2000.
- [13] Shantanu Godbole and Sunita Sarawagi. Discriminative methods for multi-labeled classification. In *Advances in Knowledge Discovery and Data Mining*, pages 22–30. Springer, 2004.
- [14] André Goffeau, BG Barrell, Howard Bussey, RW Davis, Bernard Dujon, Heinz Feldmann, Francis Galibert, JD Hoheisel, Cr Jacq, Michael Johnston, et al. Life with 6000 genes. *Science*, 274(5287):546–567, 1996.

- [15] Lesley H Greene, Tony E Lewis, Sarah Addou, Alison Cuff, Tim Dallman, Mark Dibley, Oliver Redfern, Frances Pearl, Rekha Nambudiry, Adam Reid, et al. The cath domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic acids research*, 35(suppl 1):D291–D297, 2007.
- [16] AA Hagberg, DA Schult, and M Renieris. Pygraphviz a python interface to the graphviz graph layout and visualization package.
- [17] Tobias Hamp, Rebecca Kassner, Stefan Seemayer, Esmeralda Vicedo, Christian Schaefer, Dominik Achten, Florian Auer, Ariane Boehm, Tatjana Braun, Maximilian Hecht, et al. Homology-based inference sets the bar high for protein function prediction. *BMC bioinformatics*, 14(Suppl 3):S7, 2013.
- [18] Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.
- [19] Liisa Holm and Chris Sander. Protein structure comparison by alignment of distance matrices. *Journal of molecular biology*, 233(1):123–138, 1993.
- [20] Liisa Holm and Chris Sander. Protein structure comparison by alignment of distance matrices. *Journal of molecular biology*, 233(1):123–138, 1993.
- [21] Sarah Hunter, Rolf Apweiler, Teresa K Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, Ujjwal Das, Louise Daugherty, Lاورanne Duquenne, et al. Interpro: the integrative protein signature database. *Nucleic acids research*, 37(suppl 1):D211–D215, 2009.
- [22] Vishwanath R Iyer, Michael B Eisen, Douglas T Ross, Greg Schuler, Troy Moore, Jeffrey CF Lee, Jeffrey M Trent, Louis M Staudt, James

- Hudson, Mark S Boguski, et al. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283(5398):83–87, 1999.
- [23] Rachel Kolodny, Patrice Koehl, and Michael Levitt. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *Journal of molecular biology*, 346(4):1173–1188, 2005.
- [24] Risi Imre Kondor and John Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *ICML*, volume 2, pages 315–322, 2002.
- [25] Yiannis Kourmpetis, Aalt DJ van Dijk, and Cajo JF ter Braak. Gene ontology consistent protein function prediction: the falcon algorithm applied to six eukaryotic genomes. *Algorithms for Molecular Biology*, 8(1):10, 2013.
- [26] Christina S Leslie, Eleazar Eskin, Adiel Cohen, Jason Weston, and William Stafford Noble. Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4):467–476, 2004.
- [27] Christina S Leslie, Eleazar Eskin, and William Stafford Noble. The spectrum kernel: A string kernel for svm protein classification. In *Pacific symposium on biocomputing*, volume 7, pages 566–575, 2002.
- [28] Alexey G Murzin, Steven E Brenner, Tim Hubbard, and Cyrus Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4):536–540, 1995.
- [29] Gulisong Nasierding and Abbas Z Kouzani. Comparative evaluation of multi-label classification methods. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on*, pages 679–683. IEEE, 2012.

- [30] Gaurav Pandey, Vipin Kumar, and Michael Steinbach. Computational approaches for protein function prediction: A survey. *Twin Cities: Department of Computer Science and Engineering, University of Minnesota*, 2006.
- [31] WR Pearson and DJ Lipman. Search for dna homologies was performed with the fasta program. *Proc. Natl. Acad. Sci. USA*, 85:2444–2448, 1988.
- [32] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [33] Adam Platt, Helen C Ross, Steven Hankin, and Richard J Reece. The insertion of two amino acids into a transcriptional inducer converts it into a galactokinase. *Proceedings of the National Academy of Sciences*, 97(7):3154–3159, 2000.
- [34] Shuye Pu, Jessica Wong, Brian Turner, Emerson Cho, and Shoshana J Wodak. Up-to-date catalogues of yeast protein complexes. *Nucleic acids research*, 37(3):825–831, 2009.
- [35] Predrag Radivojac, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, et al. A large-scale evaluation of computational protein function prediction. *Nature methods*, 10(3):221–227, 2013.
- [36] Claudio Saccà, Stefano Teso, Michelangelo Diligenti, and Andrea Passerini. Improved multi-level protein-protein interaction prediction

- with semantic-based regularization. *BMC bioinformatics*, 15(1):103, 2014.
- [37] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *Computational learning theory*, pages 416–426. Springer, 2001.
- [38] Roded Sharan, Igor Ulitsky, and Ron Shamir. Network-based prediction of protein function. *Molecular systems biology*, 3(1), 2007.
- [39] Hagit Shatkay, Stephen Edwards, W John Wilbur, and Mark Boguski. Genes, themes and microarrays. *ISMB2000: 2000; San Diego, California*, pages 317–328, 2000.
- [40] James Shipman, Jerry Wilson, and Charles Higgins. *An introduction to physical science*. Cengage Learning, 2012.
- [41] Christian JA Sigrist, Lorenzo Cerutti, Nicolas Hulo, Alexandre Gattiker, Laurent Falquet, Marco Pagni, Amos Bairoch, and Philipp Bucher. Prosite: a documented database using patterns and profiles as motif descriptors. *Briefings in bioinformatics*, 3(3):265–274, 2002.
- [42] Ian Sillitoe, Alison L Cuff, Benoit H Dessailly, Natalie L Dawson, Nicholas Furnham, David Lee, Jonathan G Lees, Tony E Lewis, Romain A Studer, Robert Rentzsch, et al. New functional families (funfams) in cath to improve the mapping of conserved functional sites to 3d structures. *Nucleic acids research*, page gks1211, 2012.
- [43] Paul T Spellman, Gavin Sherlock, Michael Q Zhang, Vishwanath R Iyer, Kirk Anders, Michael B Eisen, Patrick O Brown, David Botstein, and Bruce Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297, 1998.

- [44] Lawrence W Stanton, Lisa J Garrard, Deborah Damm, Brett L Garrick, Andrew Lam, Ann M Kapoun, Qiang Zheng, Andrew A Protter, George F Schreiner, and R Tyler White. Altered patterns of gene expression in response to myocardial infarction. *Circulation research*, 86(9):939–945, 2000.
- [45] Julie D Thompson, Desmond G Higgins, and Toby J Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22(22):4673–4680, 1994.
- [46] Lusheng Wang and Tao Jiang. On the complexity of multiple sequence alignment. *Journal of computational biology*, 1(4):337–348, 1994.
- [47] Kevin Y Yip, Philip M Kim, Drew McDermott, and Mark Gerstein. Multi-level learning: improving the prediction of protein, domain and residue interactions by allowing information flow between levels. *BMC bioinformatics*, 10(1):241, 2009.